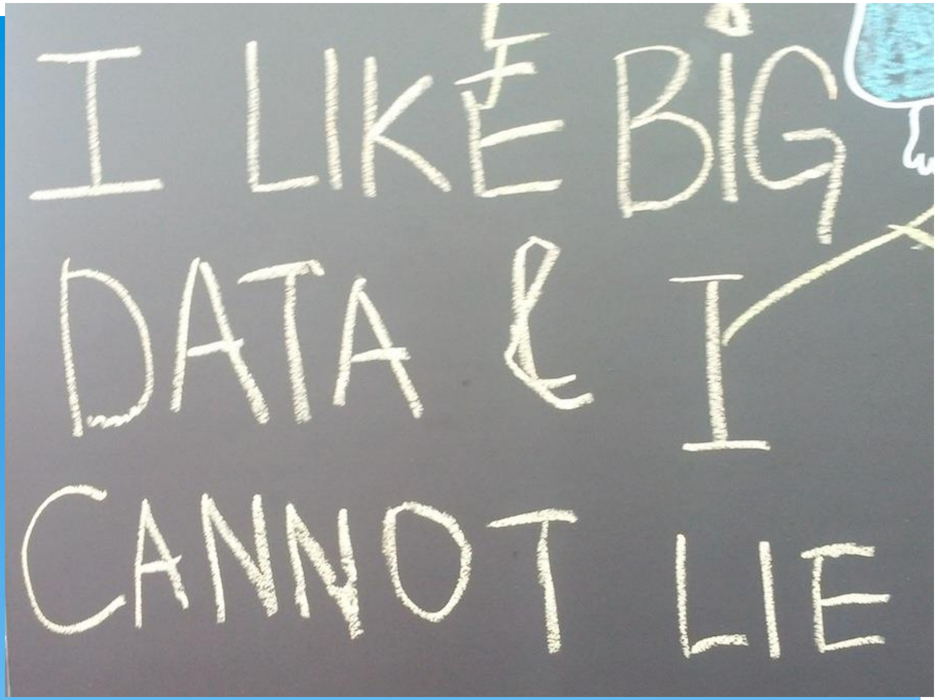
Fisheries Management Techniques FT 211

Joel Markis

Week 3

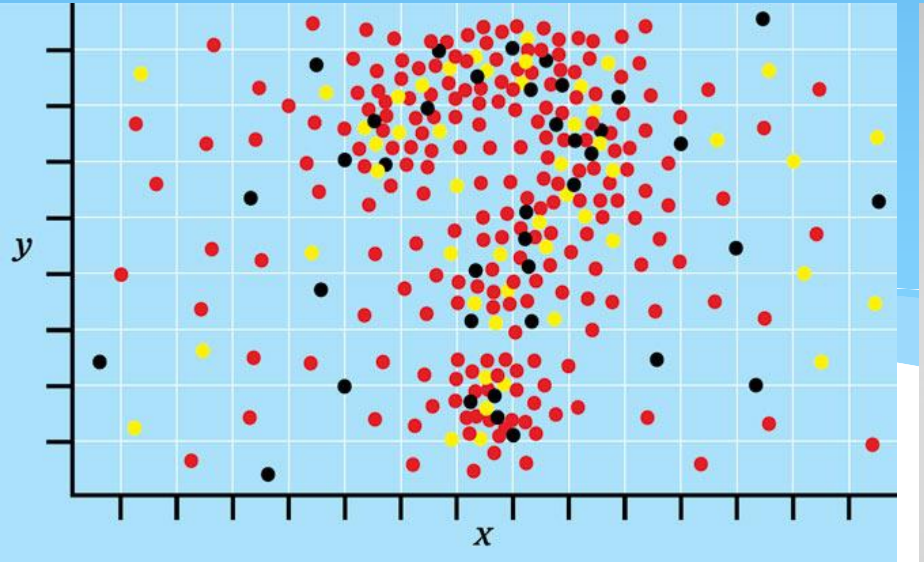
Data & Statistics

Fisheries Technology



Chapter 2

Data Management and Statistical Techniques



This Module will Contain

This Module will Contain ?? Main areas

- What are data and Statistics
- Sampling design
- Data collection in the field
- Computer management / Databases
- Data Visualization
- Overview of statistics
- Descriptive Statistics
- Inferential statistics

Student Learning Outcomes

Students will be able to:

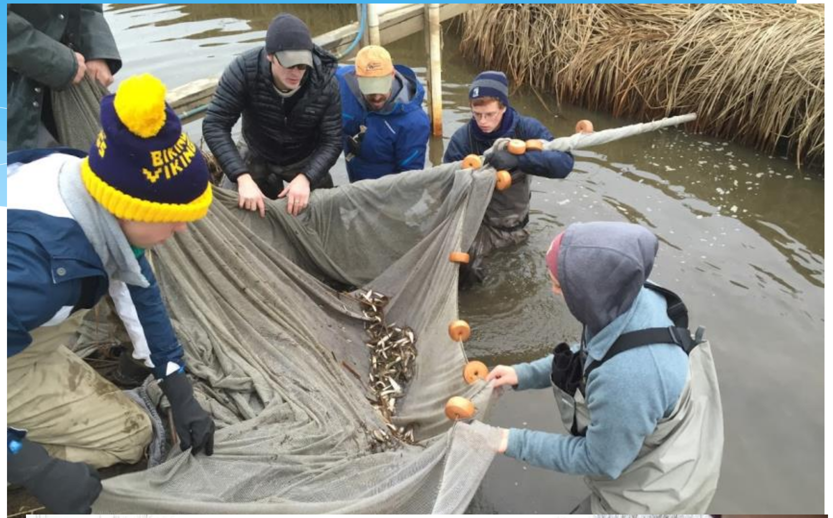
- Broadly summarize what data are how statistics can be used on data
- Overview study and sample Design
- Explain and summarize field data collection techniques
- Outline computer based and database handling of fisheries data
- Summarize types of data visualization
- Demonstrate an understanding of general statistical concepts
- Define descriptive statistical techniques
- Summarize inferential statistical concepts

Fisheries Techniques Field Course

Dates

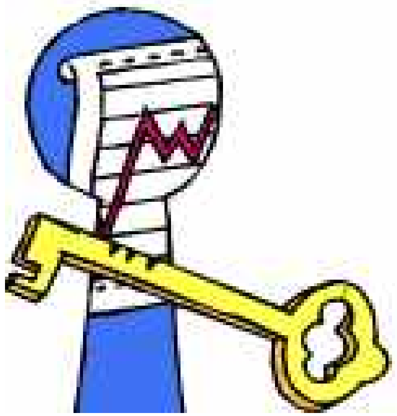
- Ketchikan – April 15 – 17
- Kodiak – April 22 – 24

Sign Up if You haven't Already!



Data and Statistics in Fisheries

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



What do we collect?

Data

What are data?

Values of quantitative or qualitative variables belonging to a set

Special Note: Data is the plural form of datum

- so one says, "The data are..."
- These Data
- not "The data is...“

Statistics

Statistics - is the study of the collection, analysis, interpretation, presentation, and organization of data.

http://www.ted.com/talks/arthur_benjamin_s_formula_for_changing_math_education

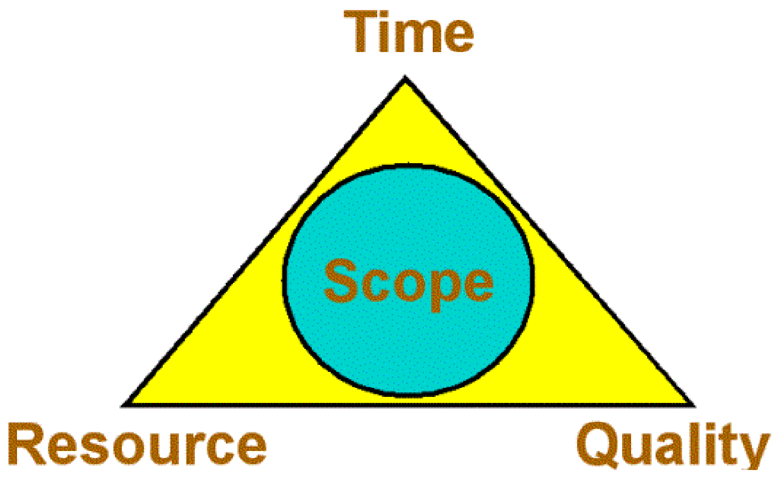
- Analyzing and Interpreting data
- Inferences from a sample to the population

Statistician

- Likes figures but lacks the social skills to be an accountant
- https://www.youtube.com/watch?v=IUK6zjtUj0o
- "There are Lies, damned lies, and statistics“
  - British Prime Minister Benjamin Disraeli but **Mark Twain**

Audience, Scope, and Limitations

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



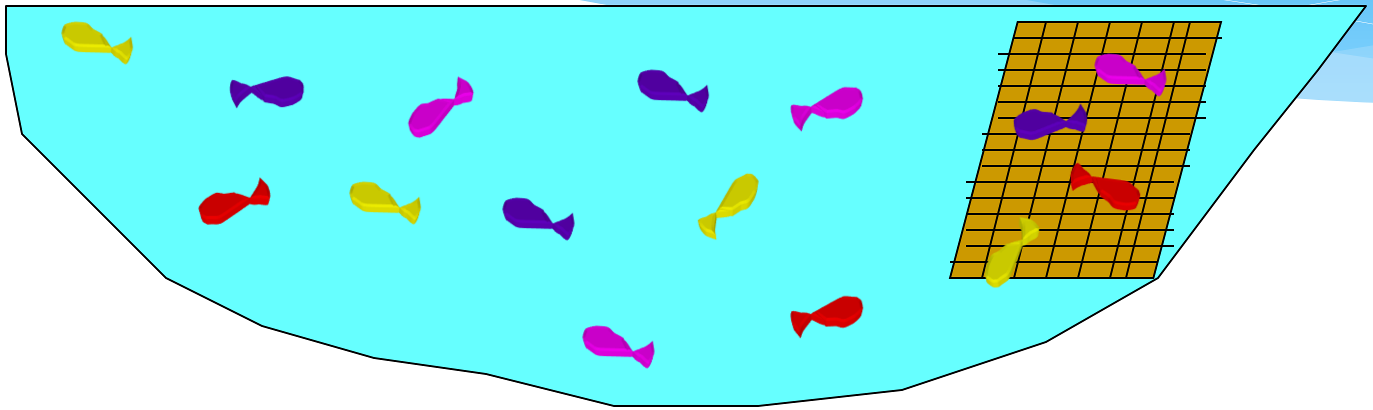
Self Check 1

- Generally speaking statistics involves Analyzing and Interpreting data
  - True
  - False

- Who said “There are Lies, damned lies, and statistics”
  - Earnest Hemingway
  - **Mark Twain**
  - William Faulkner
  - John Steinbeck
  - F. Scott Fitzgerald

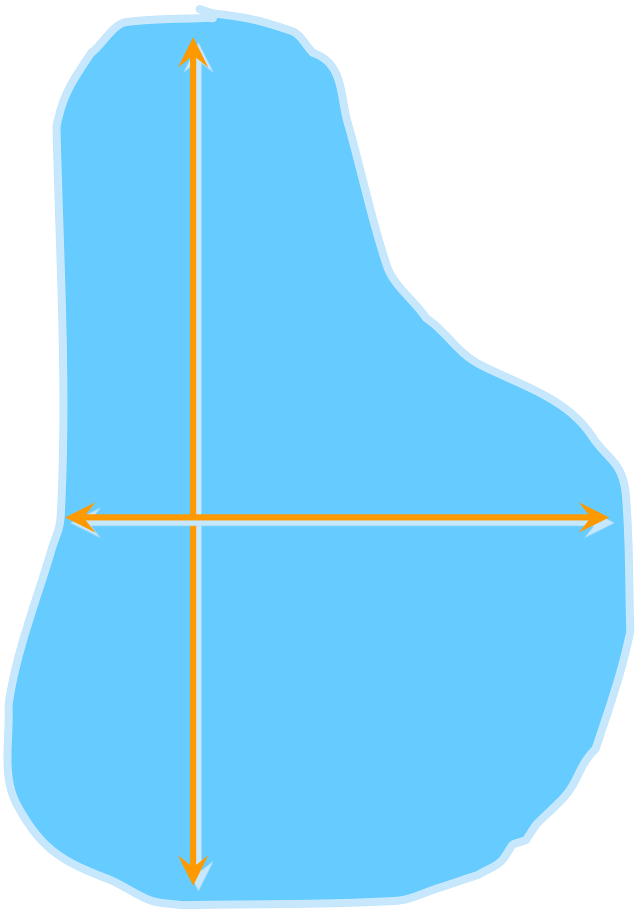
Collecting Data and Statistics

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

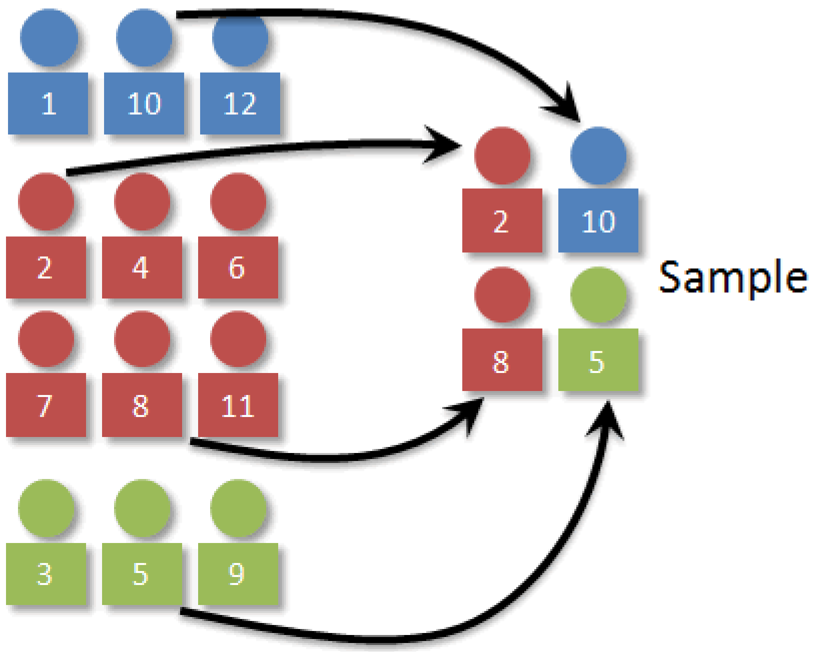
- Every member of the population has equal opportunity to be sampled
- With or without replacement
  - Sleepy fish will be easier to catch
- Random number table

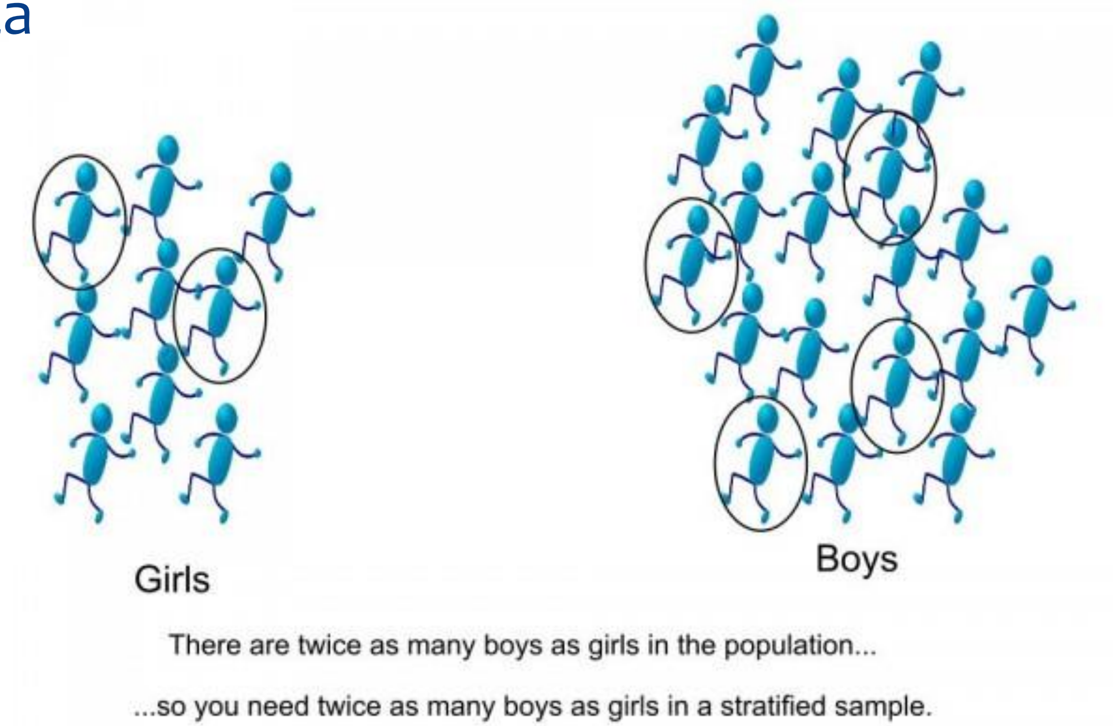
Part of a Table of Random Numbers

61424	20419	86546	00517
90222	27993	04952	66762
50349	71146	97668	86523
85676	10005	08216	25906
02429	19761	15370	43882
90519	61988	40164	15815
20631	88967	19660	89624
89990	78733	16447	27932

Stratified random sample

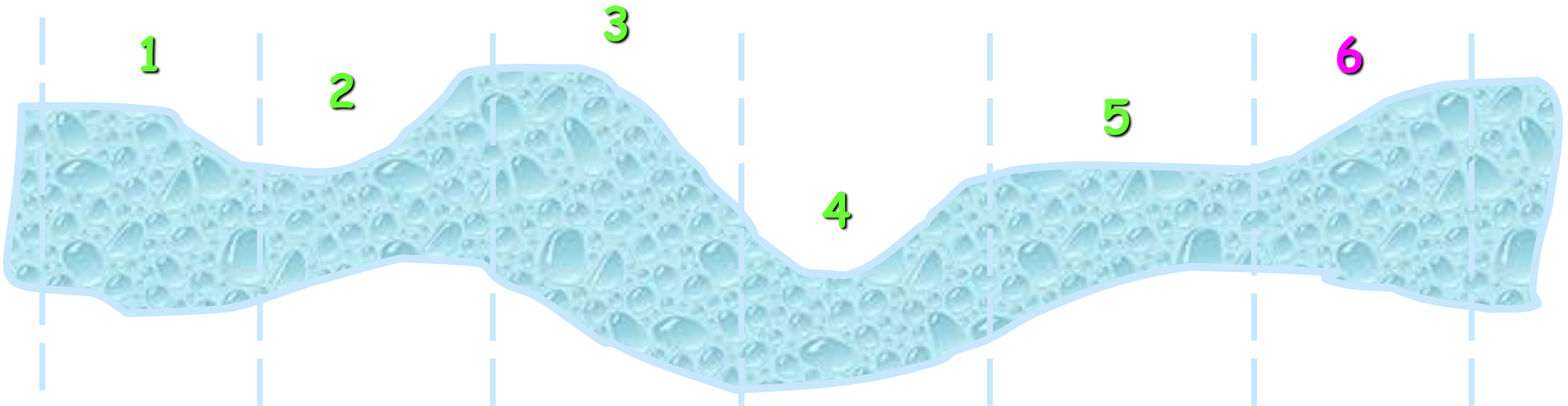
- Divide population into Strata
- Randomly sample within strata





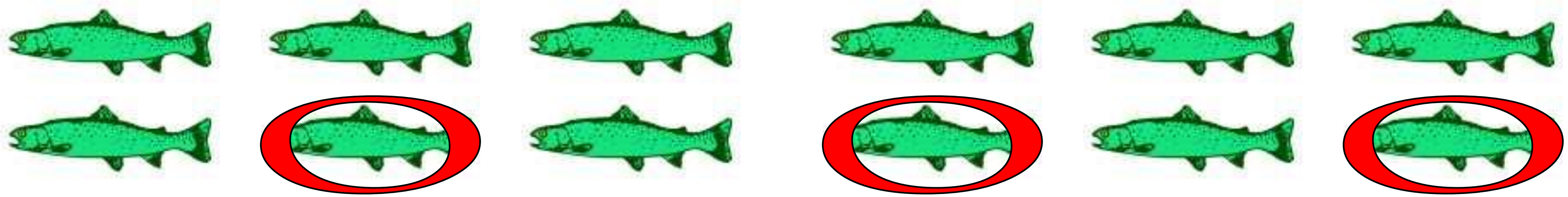
Cluster sampling

- Determine sampling sites
- Choose a site randomly
- Take all the samples from a single site



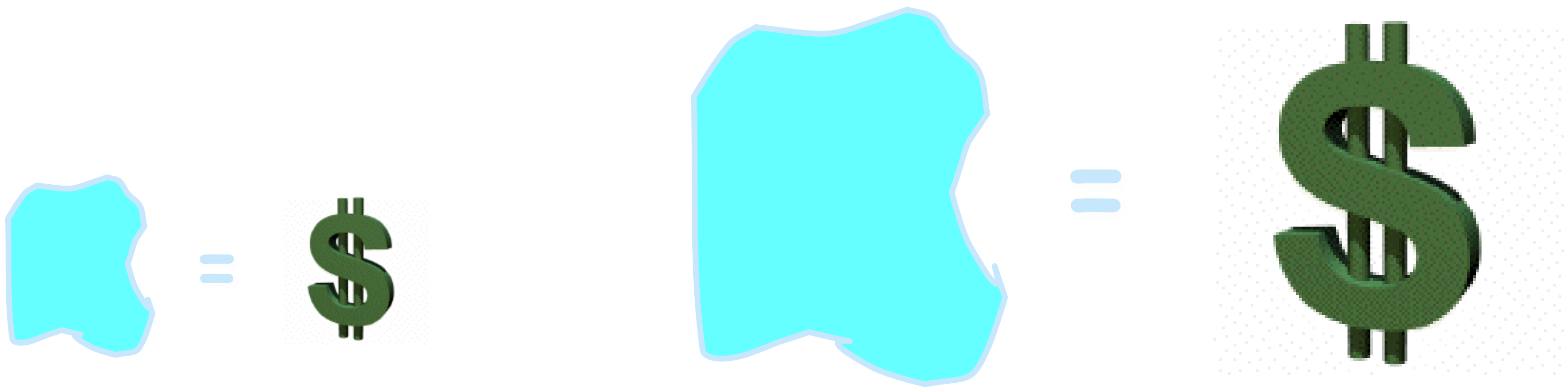
Systematic sampling

- 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



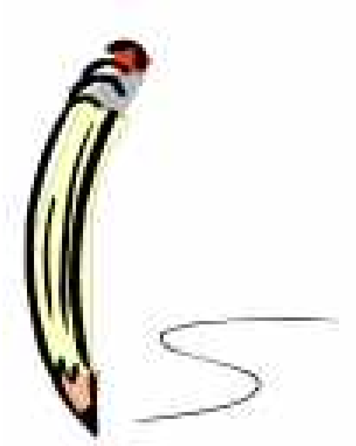
Self Check 2

- How can one generate a random number
  - Use a random number generator
  - Roil a die
  - Flip a coin
  - **All of the above**

- Dividing the population into strata and then randomly selecting a sample within strata is an example of
  - Simple random sampling
  - **Stratified random sampling**
  - Systematic sampling
  - Cluster sampling

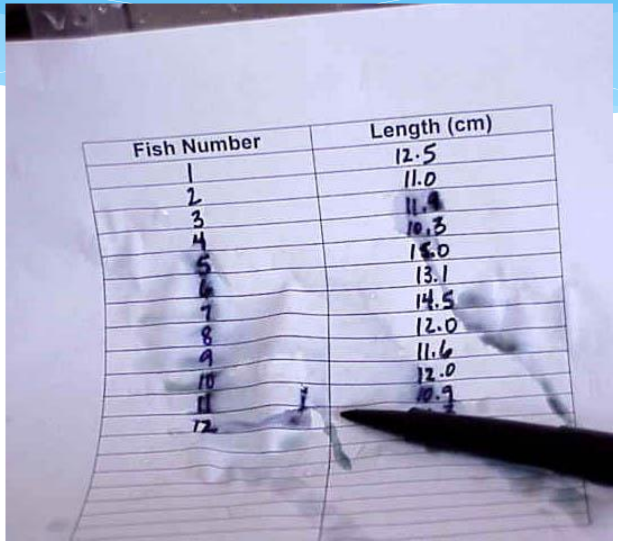
Data Handling and Database Management

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



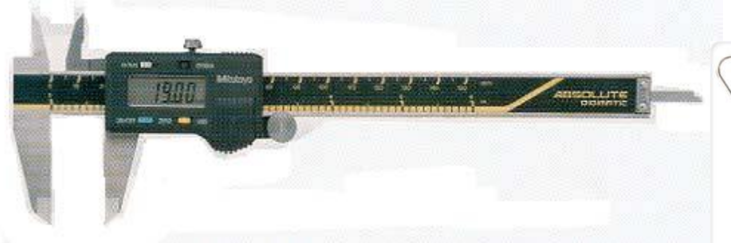
Field data sheets are standardized by study

- Print on waterproof paper
- Write with pencil, ink will run
- Write legibly, you may not be one reading
- Copy or input data sheets asap
  - Easier to resolve discrepancies when its fresh

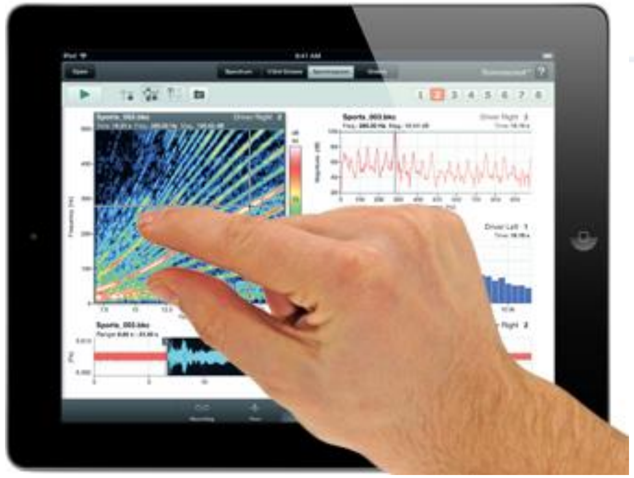


When possible, make use of new technology

- Electronic measuring boards
- Digital calipers



- iPad and dataloggers
- Check to be sure data are being recorded



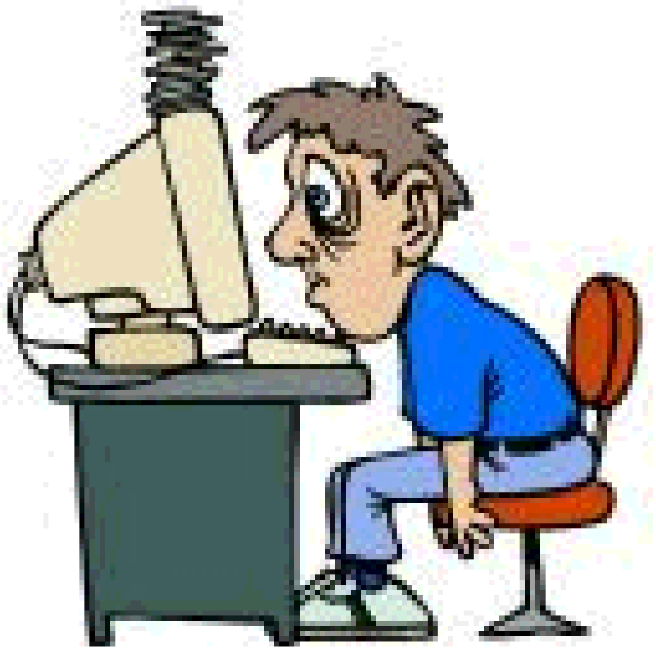
Self Check 3

- Which is best for recording written information in the field?
  - **Pencil**
  - Pen
  - Sharpie

- Electronic measuring boards and digital calipers are both examples of ways to reduce writing and data errors in data collection in the field
  - **True**
  - False

Data Management

- Most Organizations use databases. So…
- Biologists need to understand databases
- Also how to enter and retrieve data
- Database manager



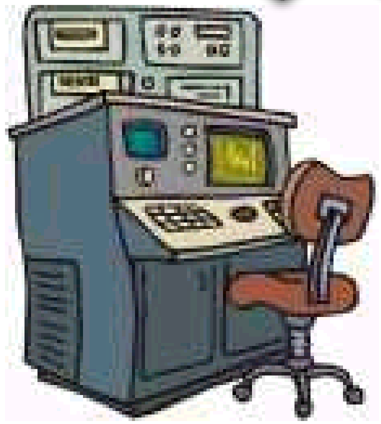
Databases are

- Repositories of information
- Logically organized
- Facilitate retrieval of specific information
- Provide for customized output reports
- Relational

Examples of databases include

- PC
  - Access
  - dBase IV
  - Paradox
  - Double Helix

- Mainframes
  - Oracle

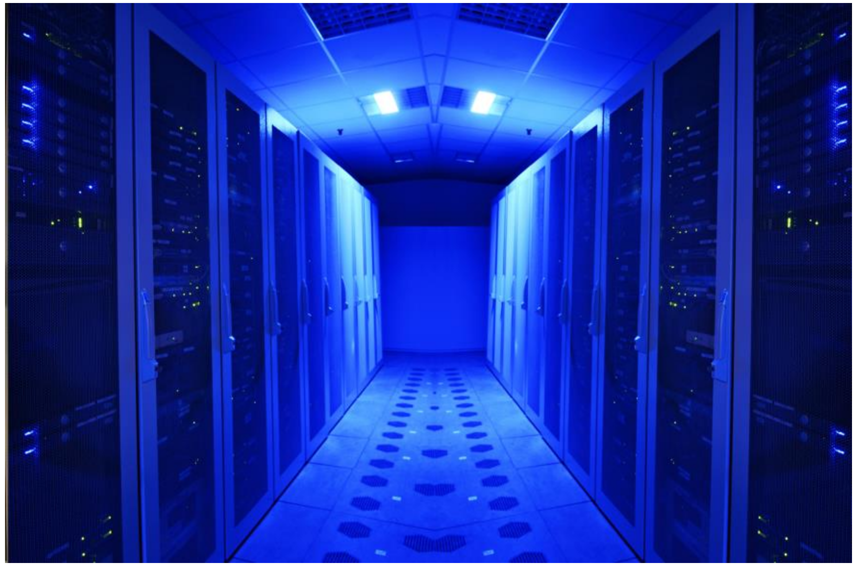


Storage Considerations

- ALWAYS MAKE BACKUPS
  - daily, weekly, monthly
- CD-ROMs may degrade after 30 years
- Technology becomes obsolete (5 1/4" floppies)
- Most organizations have network drives
- RAID Storage
- Cloud







Quality Control

QC

- What quality control exists?
  - There needs to be some!
    - Number in your pants, factory line
- Are data within believable ranges?
  - Sorting is Huge
  - NERRS Stories
  - USFS Forest Inventory
- check printouts by hand
- use two people to proofread

Self Check 3

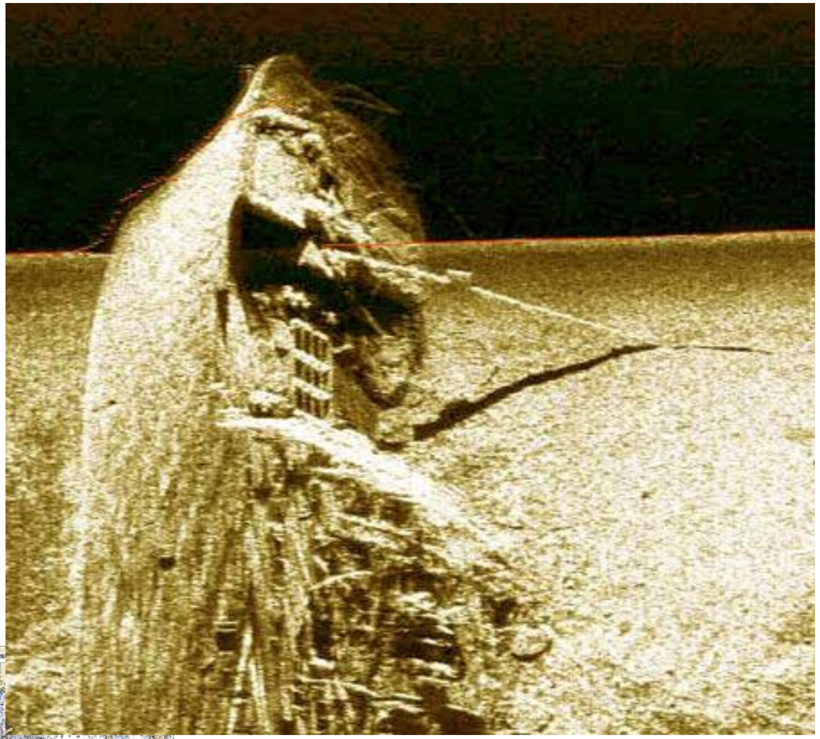
- The paper number in the pocket of a new pair of pants is an example of what?
  - Pants Database
  - Sizing Information
  - Quality Control
  - None of the above

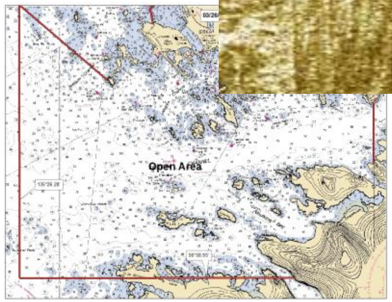
- Click on the icon that is **NOT** a type of database software

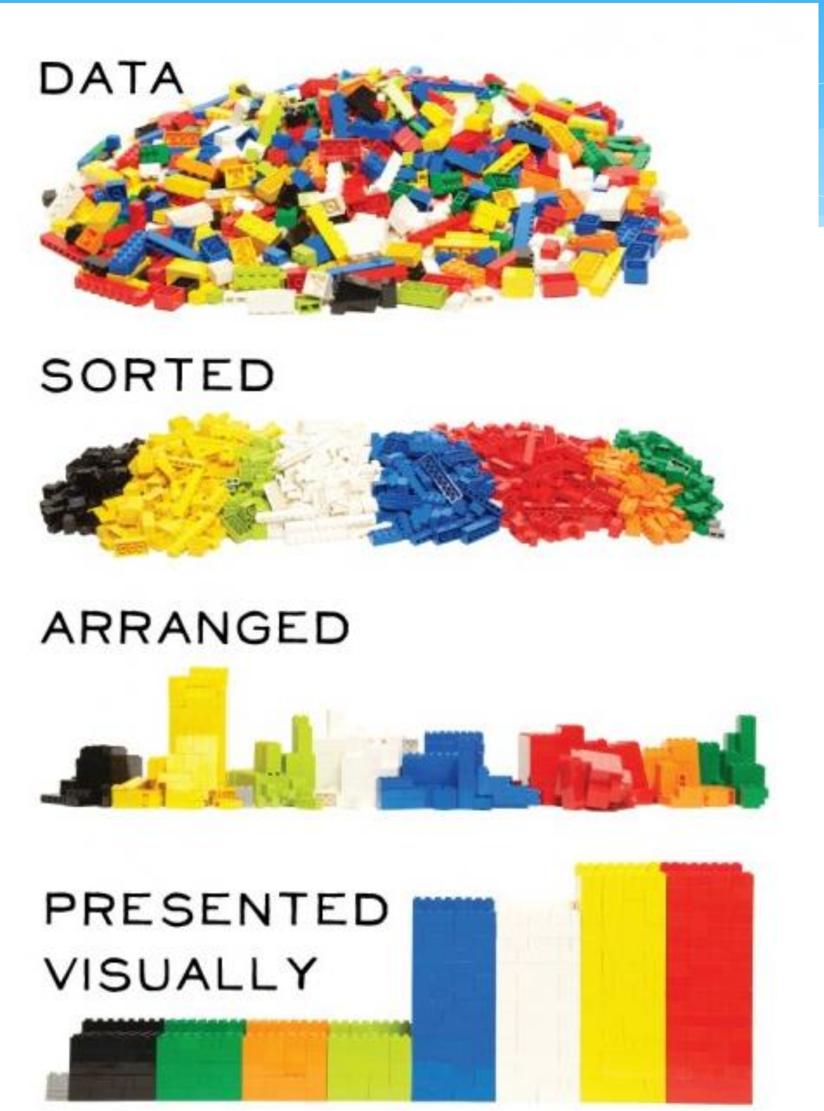
Break

Data Visualization (i.e. graphs)

Visualization is so important





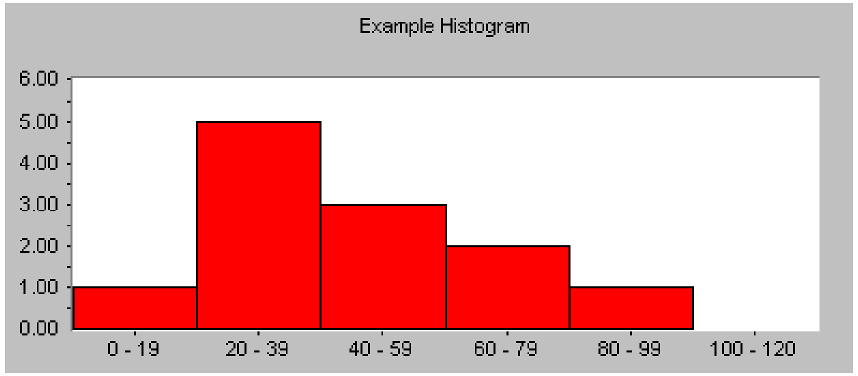


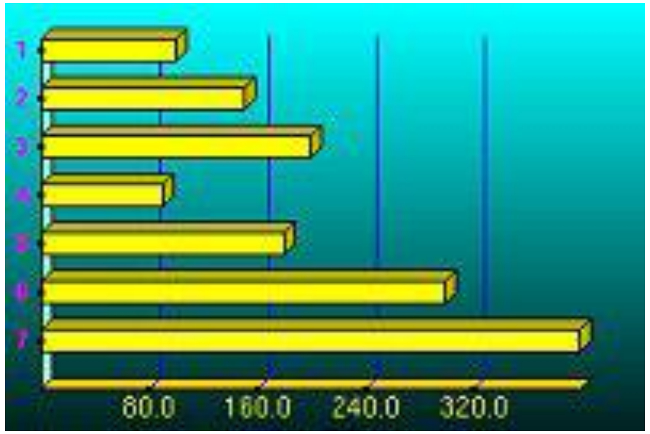
Data Visualization (i.e. graphs)

- Depict ALL data
- Picture worth 1000 numbers
  - pie chart
  - bar chart
  - histogram (vertical or horizontal)
  - scatter plot
  - line graph

Histograms and Bar Charts

- Histogram
  - Graphical representation of data
  - For continuous data
  - Length-frequency data
  - Watch out for bin size bias
- Bar Chart
  - For category data
  - Spaces between





Pie Chart

- Also for category data
- Like diet components
- Size of slice equals relative contribution

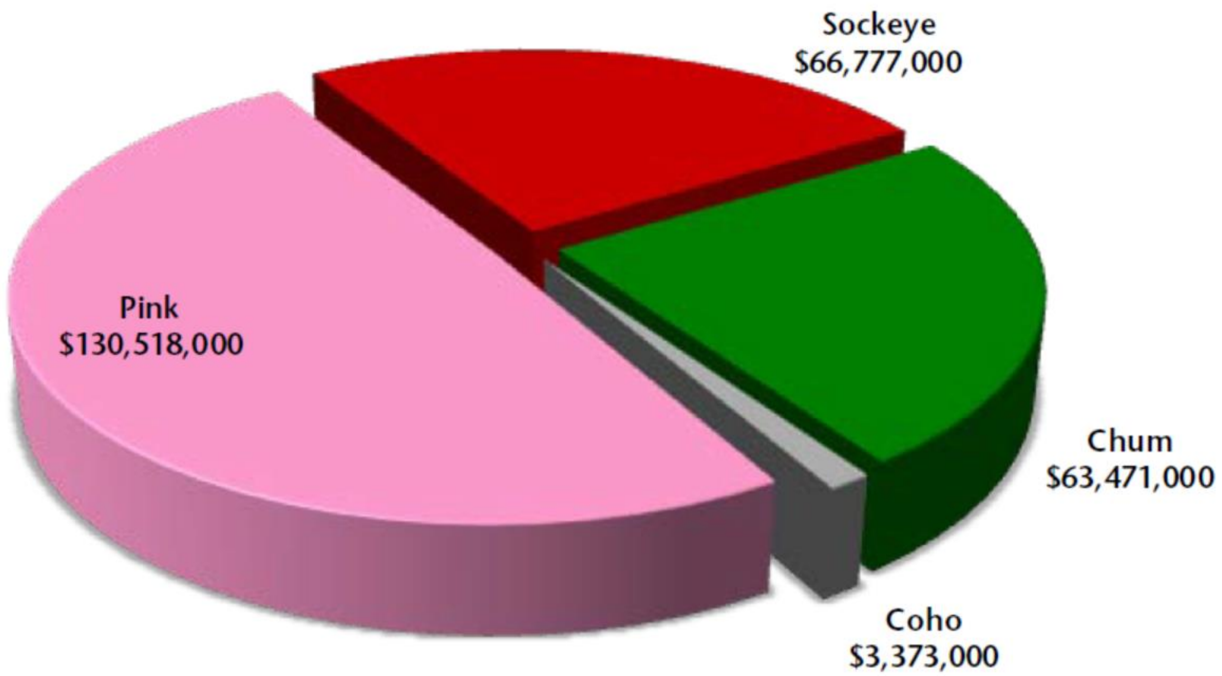
130518000 – Pink

66777000 – Sockeye

63471000 – Chum

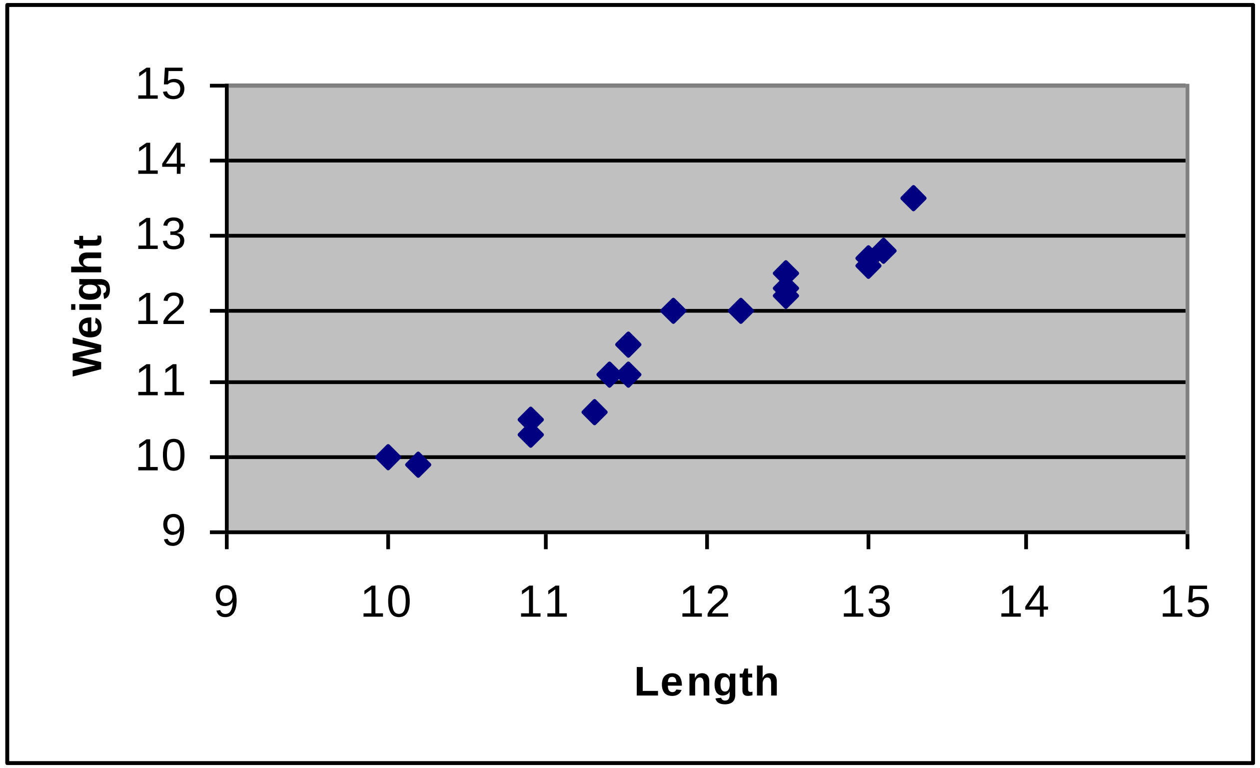
3373000 – Coho

Figure 3.4: Ex-Vessel Value of PWSAC Salmon by Species, 2007-2011 Total

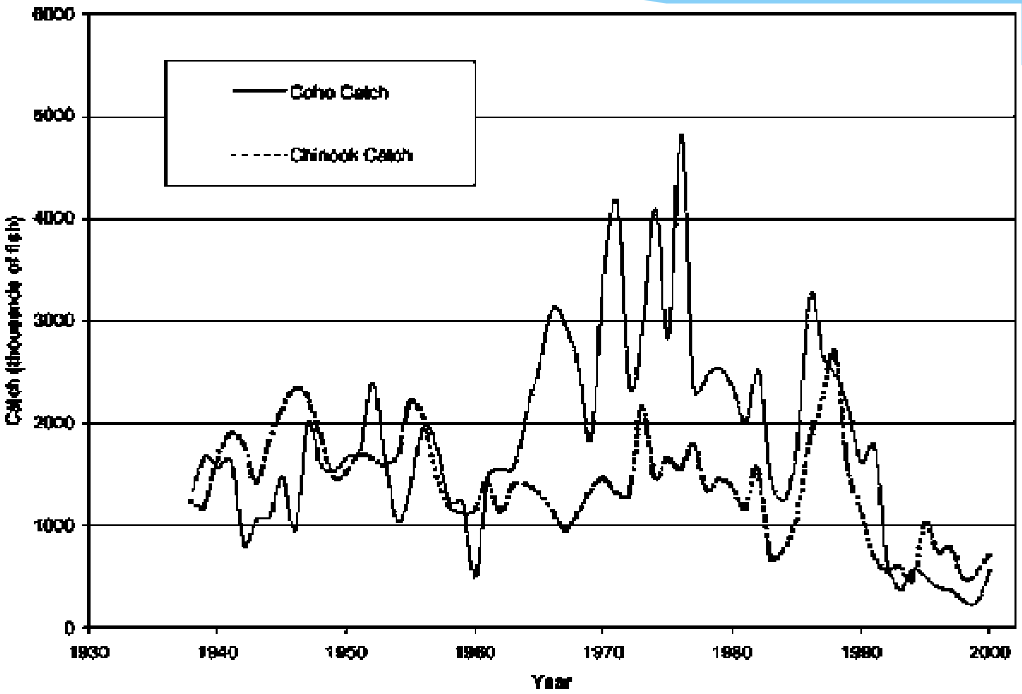


Scatter Plots

- 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

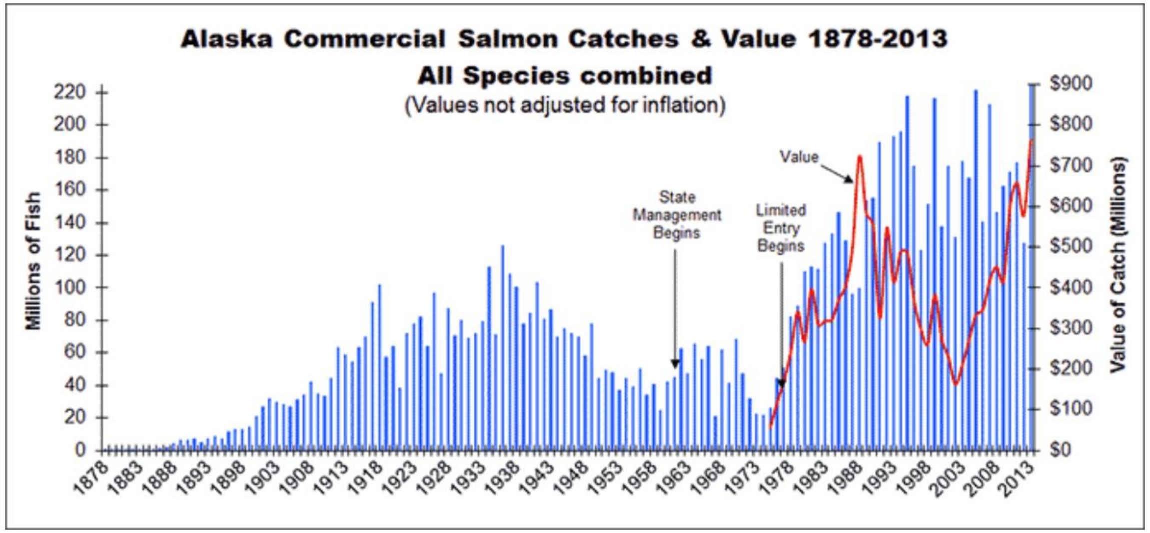


Line Graphs



time

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

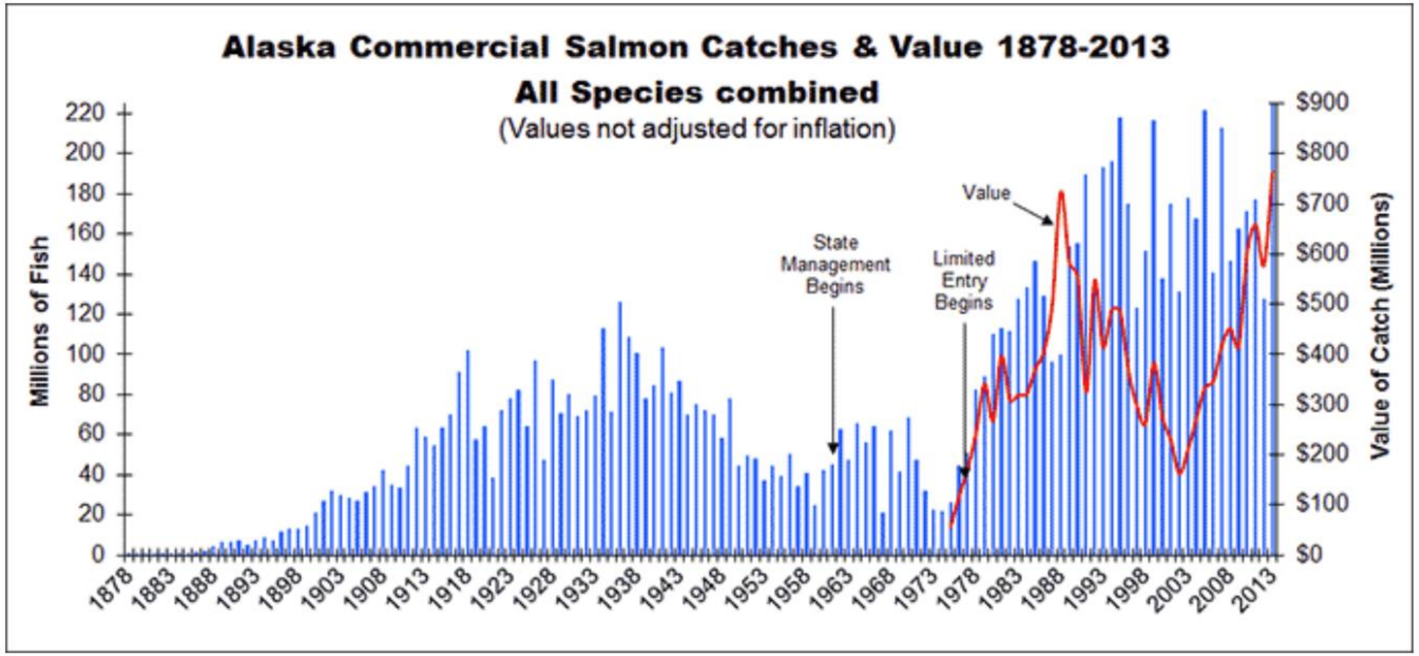


Alaska Commercial Salmon Harvests and Exvessel Values

Source: ADF&G, October 2015

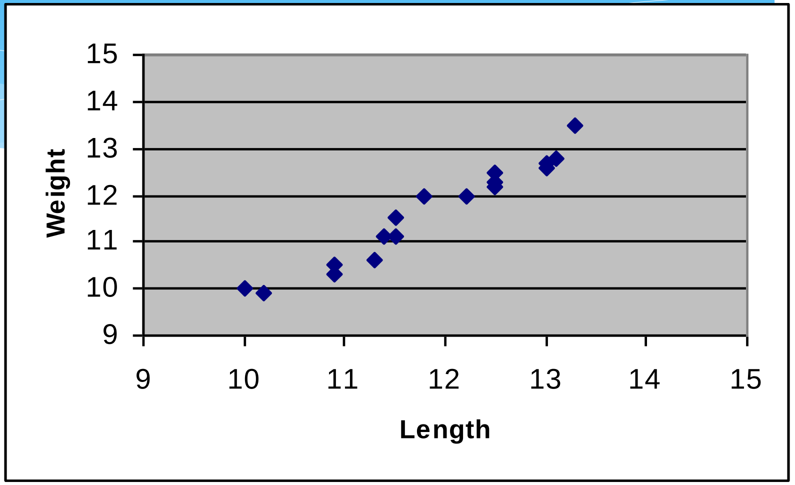
2015 Alaska Commercial Salmon Harvests and Exvessel Values					
Species	**Avg. Wt. (pounds)**	**Avg. Price per Pound**	**Number of Fish (thousands)**	**Lbs. of Fish (thousands)**	**Est. Value US$ (thousands)**
Southeast					
Chinook	10.06	$3.81	307	3,085	$11,751
Sockeye	4.36	$1.09	1,389	6,054	$6,598
Coho	5.88	$0.78	1,876	11,030	$8,604
Pink	3.84	$0.20	34,089	130,900	$26,180
Chum	8.46	$0.50	8,559	72,407	$36,204
Totals			46,218	223,473	$89,335
Prince William Sound					
Chinook	16.42	$5.65	24	388	$2,189
Sockeye	5.35	$2.01	3,210	17,183	$34,593
Coho	7.43	$0.66	198	1,469	$966
Pink	3.38	$0.22	98,254	332,085	$71,913
Chum	5.38	$0.61	2,544	13,679	$8,331
Totals			104,229	364,802	$117,990

VS

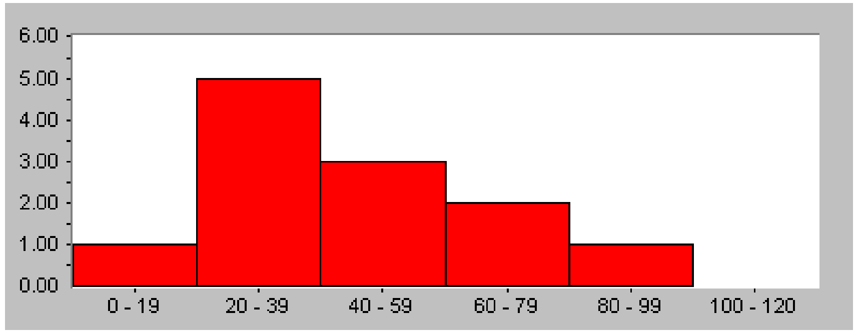


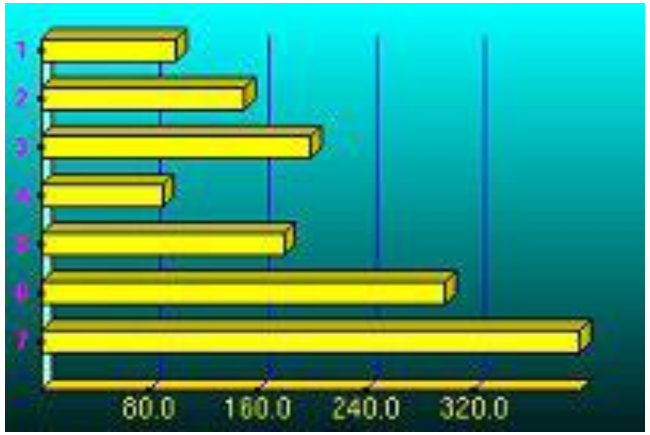
Self Check 4

- What type of data visualization is depicted above
  - Pie chart
  - Bar chart
  - Histogram
  - **Scatter plot**
  - Line graph



- A ________ is for categorical data
  - Bar chart
  - Histogram





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: age of each fish

Fish	Length	Weight	Age
1			
2			
3			
4			
5			

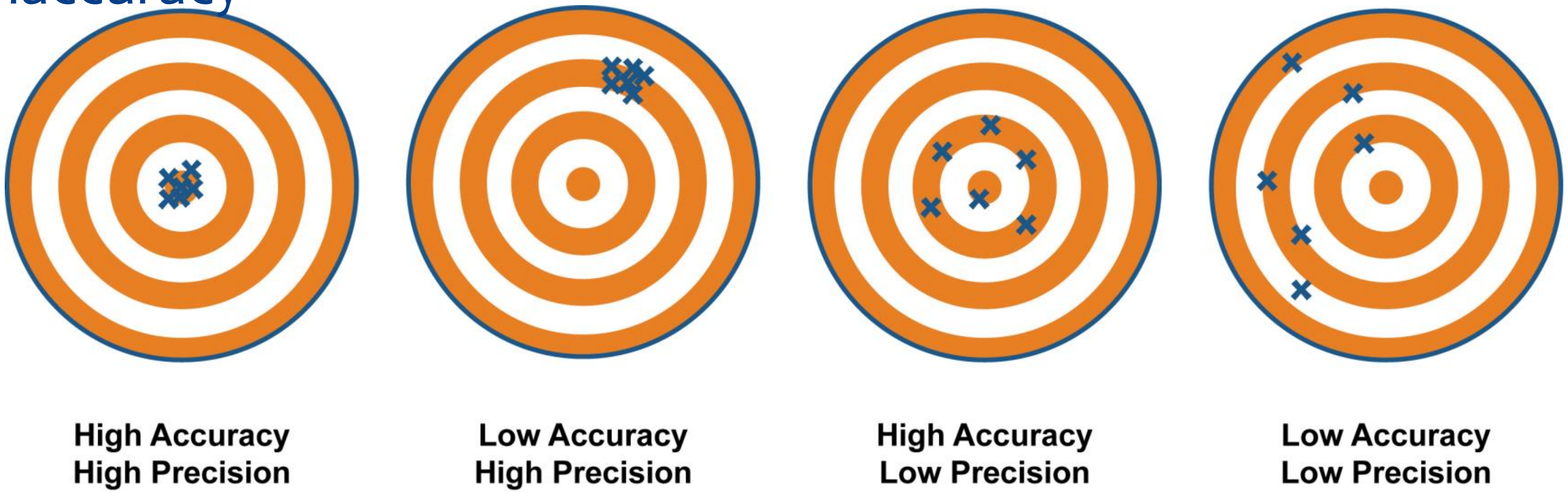
Qualitative and Quantitative data

- Qualitative = category data
  - nominal (sex, species) cannot order
  - ordinal (ranked data, house number) can order

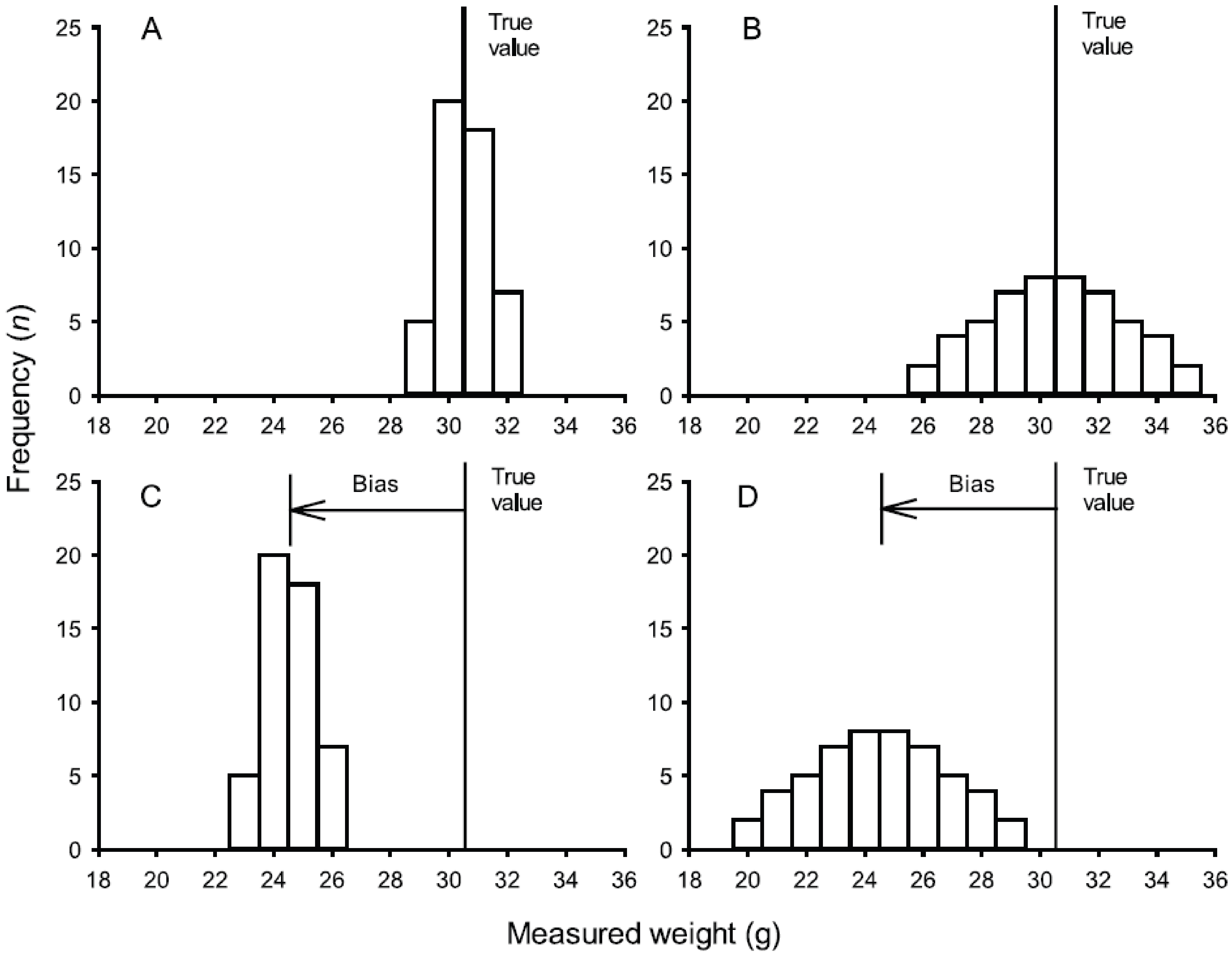
- Quantitative = numerical data
  - discrete (integers example: age, count)
  - continuous (not integers example: length, temp, time)
    - Can assume an infinite number of values between any two

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

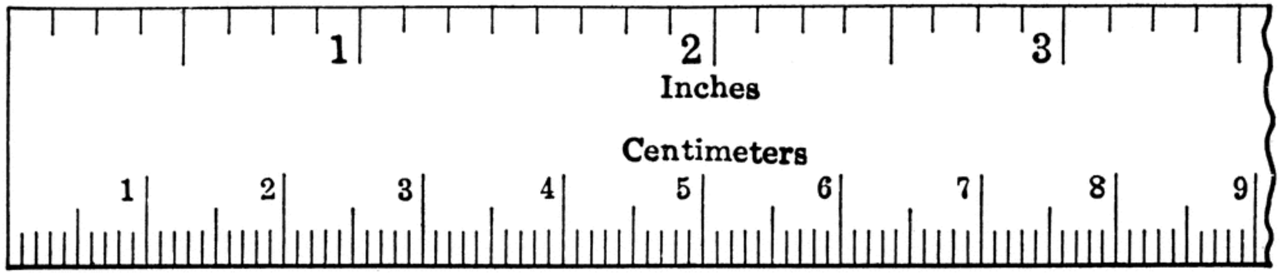


- Precise
- Accurate
- Biased



Significant digits

- Can't be more than the level of your measurement!



- Minimum accuracy = range / 30
- Maximum accuracy = range/300
  - Fish lengths 21.362 – 51.482 – Range = 30.120
    - Minimum – 30/30 = 1
      - So 1cm
    - Maximum – 30/300 = 0.1
      - So .1cm

3.14159562

Self Check 5

- The above represents which type of data variable
  - **Row**
  - Column

Fish	Length	Weight	Age
1			
2			
3			
4			
5			

- Precision represents how close a pattern is to the center of bull's eye, closer means more precise
  - True
  - **False**

Statistics

- Falls into 2 categories
- Descriptive – Collection, organization, summarization & Presentation of data
- Inferential – Generalizing from small to large, estimation, hypothesis testing, variable relationships, making predictions

Statistics

- Analyzing and Interpreting data
- Inferences from a sample to the population
  - If samples are selected accordingly - represent population
  - n = Sample Size

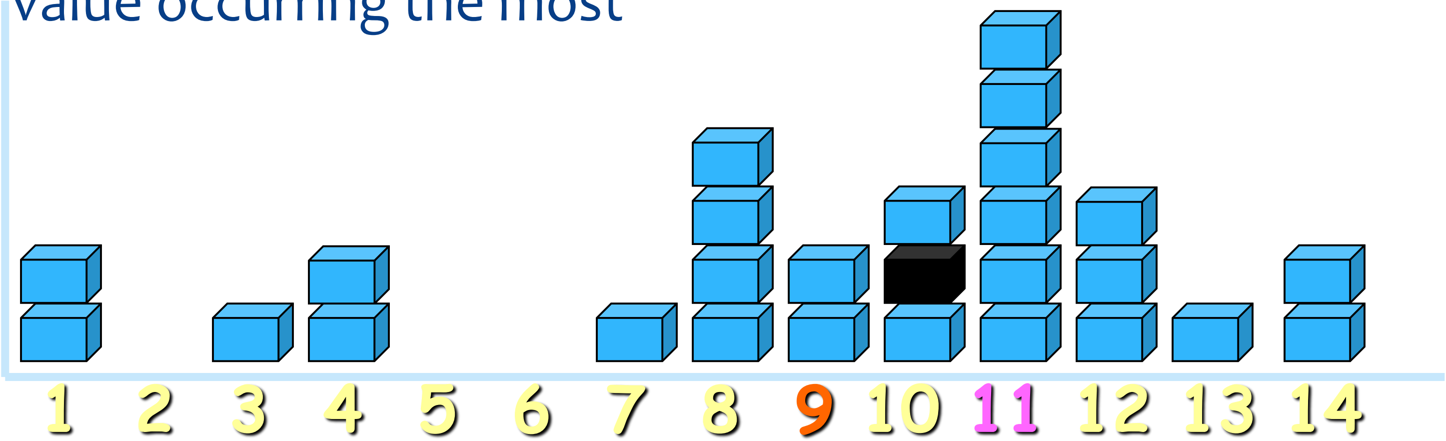
Statistics

1. Describe data
2. How spread out the data are

Descriptive Statistics

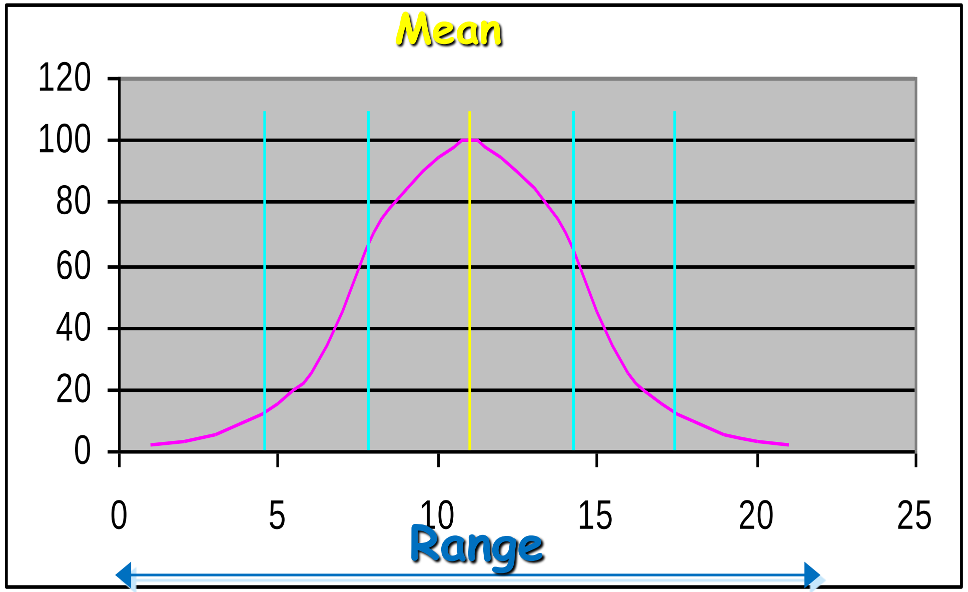
- We have data, now we want to describe this data
- Summarize lots of measurements with one number (or a few)
- Measures of central tendency
  - mean = arithmetic average
  - median = middle value
  - mode = value occurring the most

$$\bar{x} = \frac{\Sigma x}{n}$$

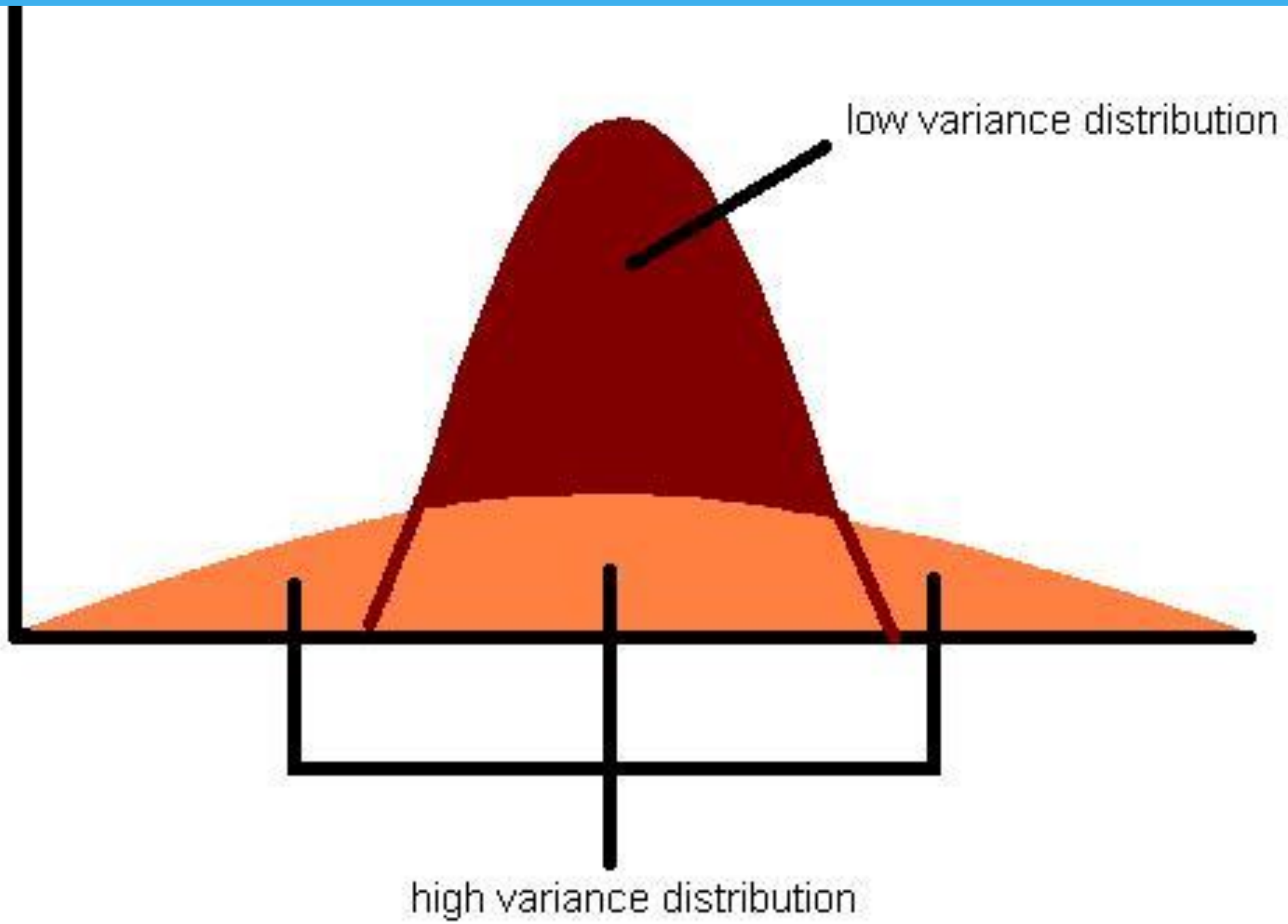


Descriptive Statistics (cont.)

- Measures of **Dispersion**
  - **Range** = max - min value
  - **Variance** = sum of squared deviations from sample mean
    - How much the data varies
- **Standard deviation** (SD)
  - square root of variance
- **Standard error** of mean (SE)
  - standard deviation / root of sample size

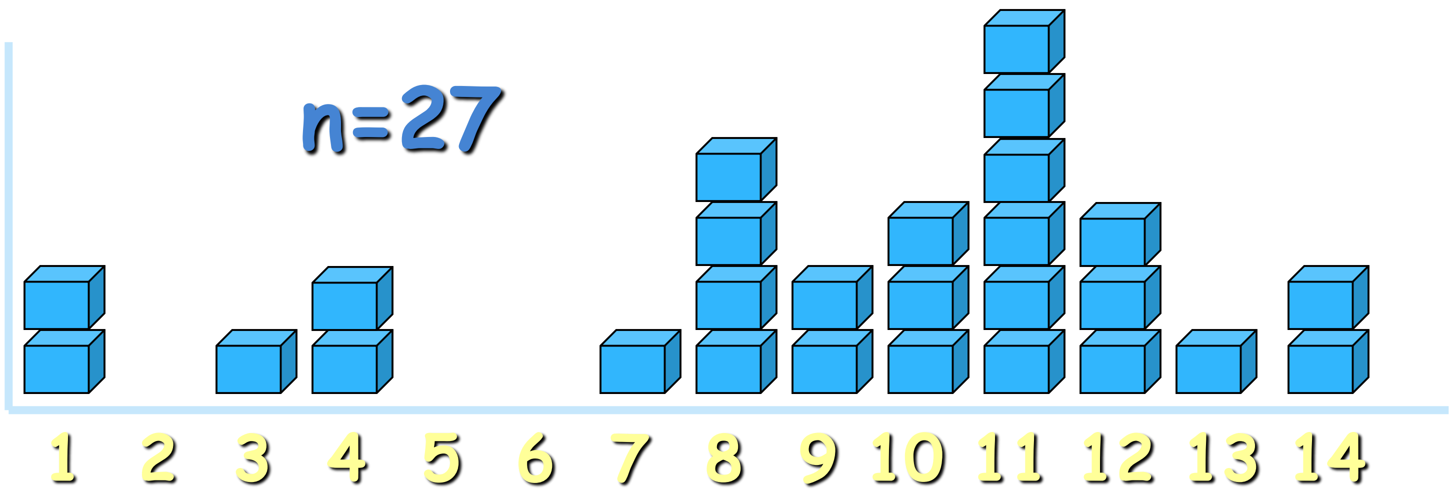


Variance



Degrees of Freedom

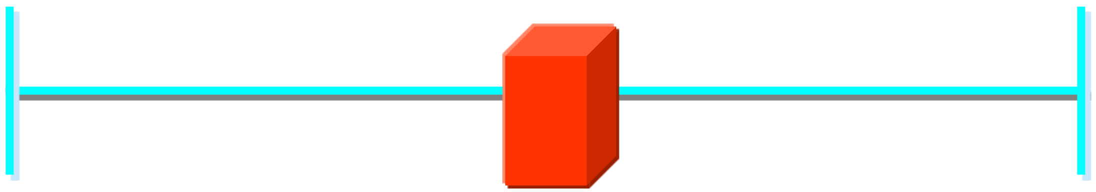
- Number of **independent observations** in data set
- **n-1** where n = number of observations
- increased degrees of freedom reduces variance



Confidence Intervals

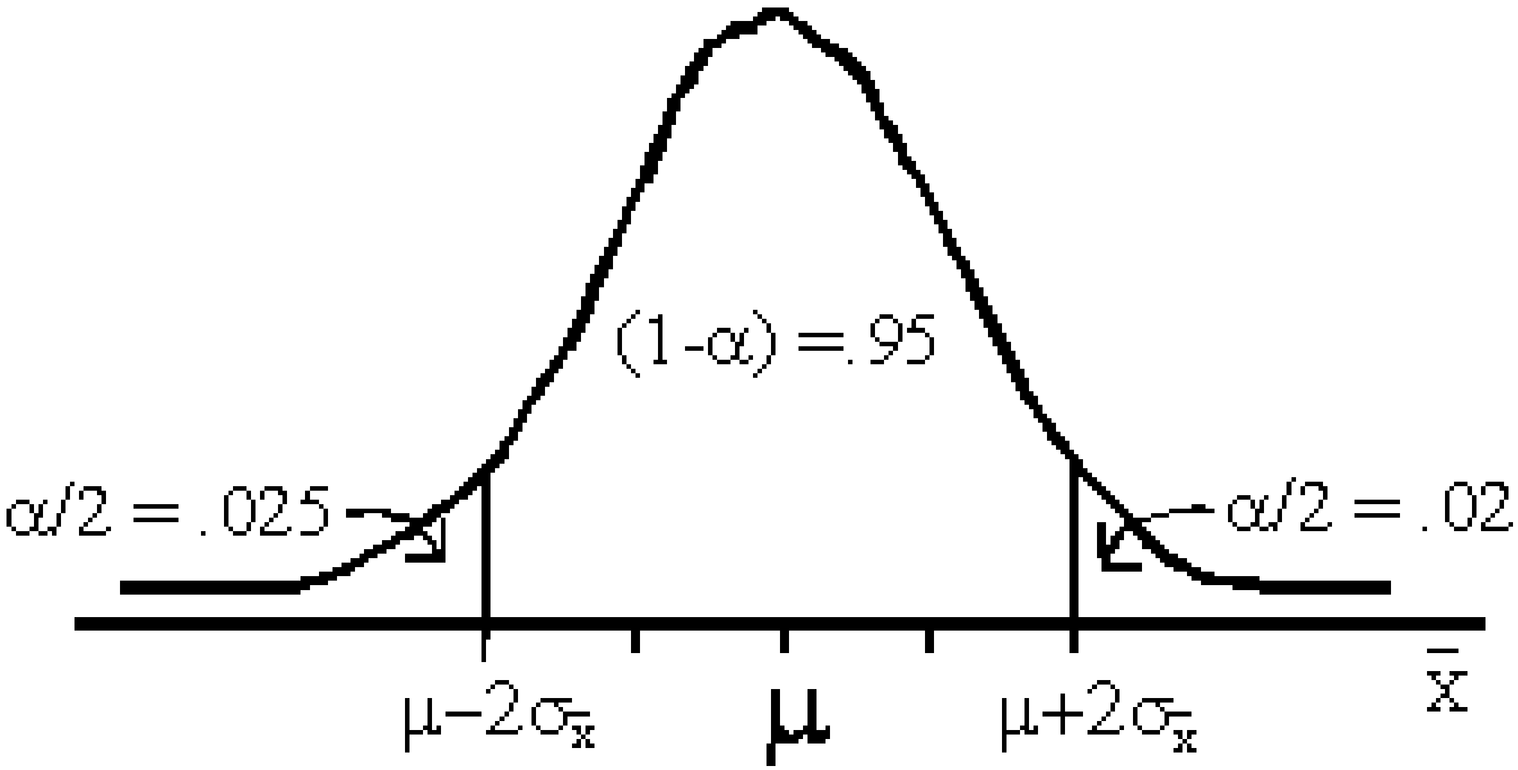
- Sample average rarely equals population mean
- Express estimate as a range of values
- Average plus/minus Student's t (n-1 df) times standard error of mean

$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$



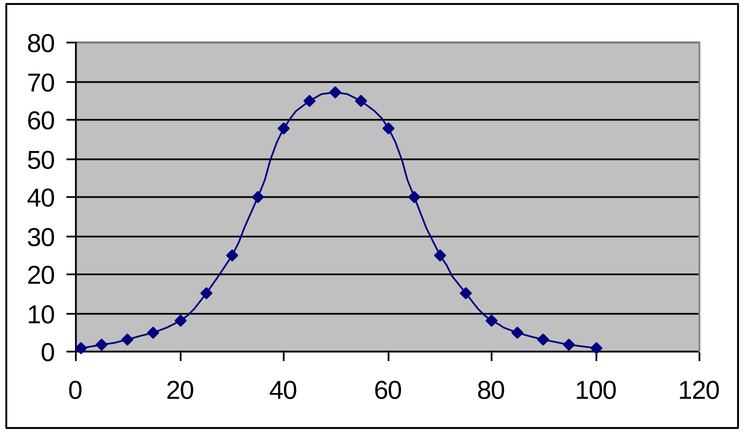
Confidence Intervals

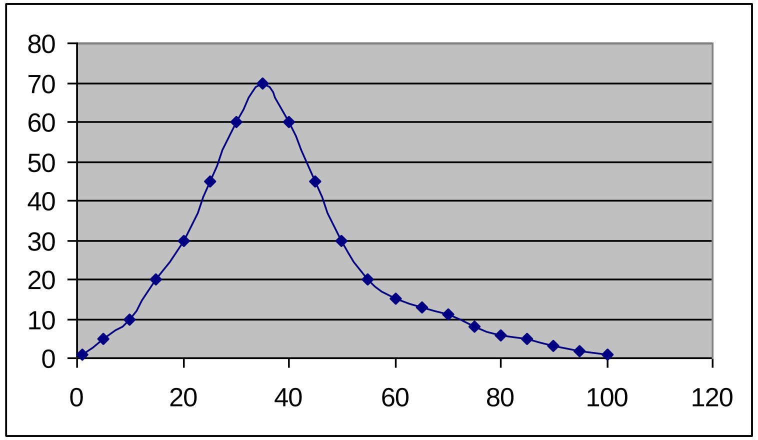
The 95% confidence interval for μ

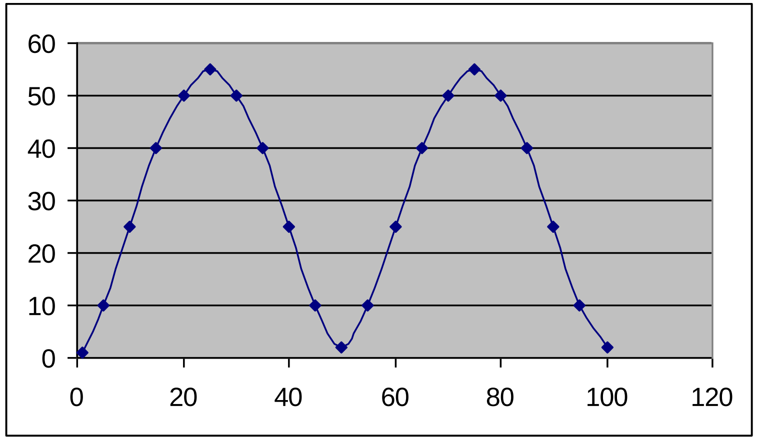


Distributions

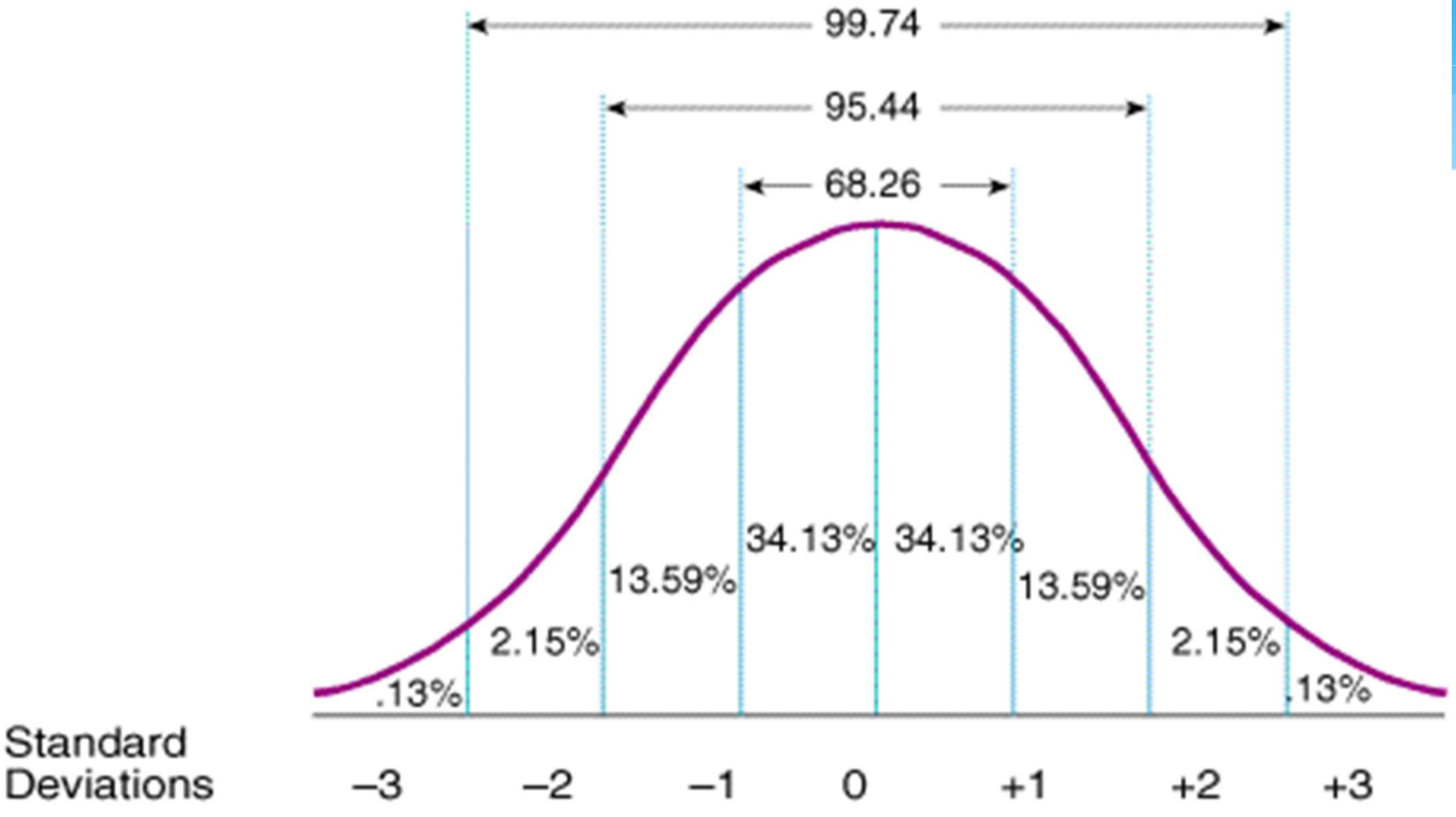
- Normal - bell shaped curve
- Skewed - data clumped to right or left
- Bimodal - two peaks in the range of data







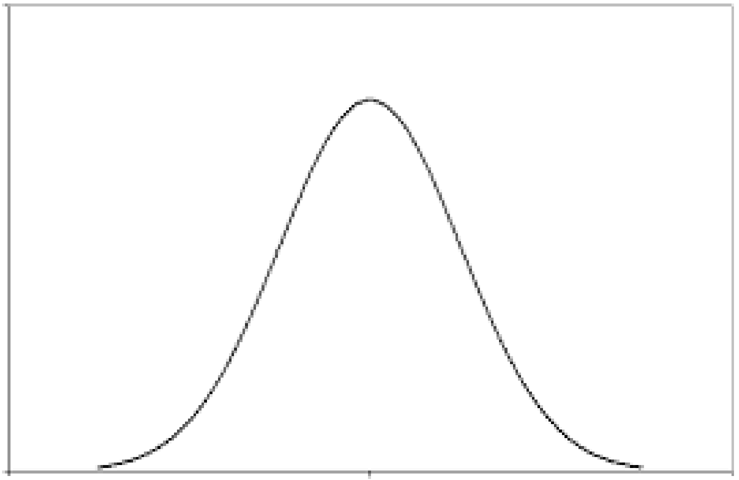
Normal Distribution



Self Check 6

- In general descriptive statistics fall into two categories, measures of the 'Central tendency' and Measures of dispersion
  - True
  - False

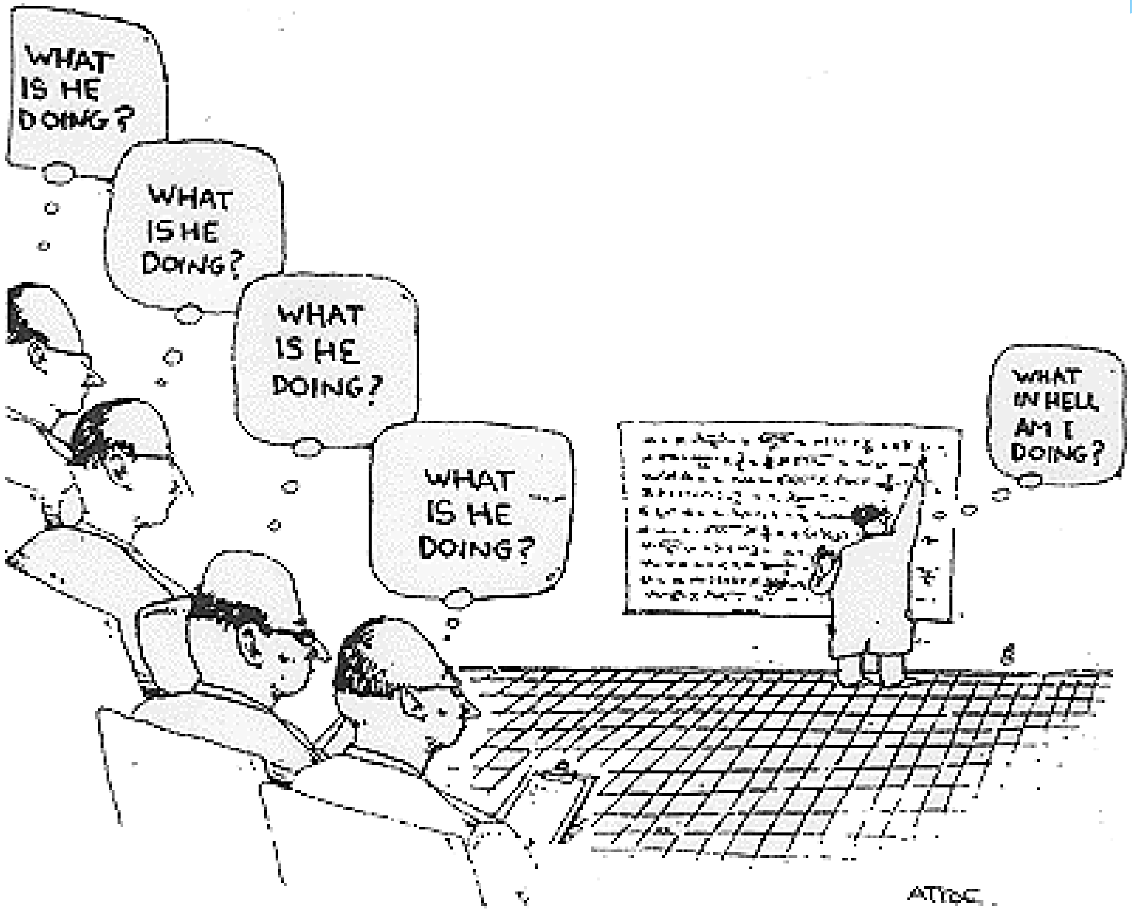
- The above figure represents what kind of data distribution
  - **Normal**
  - Sigmodal
  - Skewed
  - Bimodal



Inferential Statistics and Hypothesis Testing

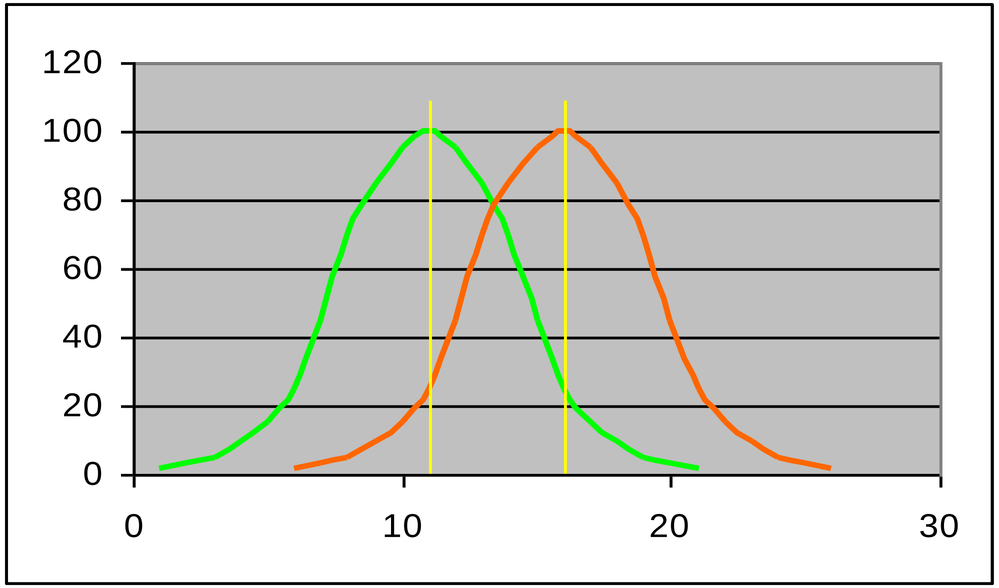
- What can we infer about the data
- Are two variables related?

- Are two groups of fish different?



Inferential Statistics and Hypothesis Testing

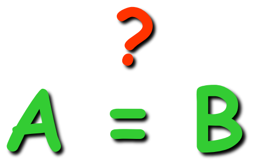
- Null hypothesis… no difference in pop means
- Two-sided alternative hypothesis… yes difference in pop means
- One-sided alternative hypothesis… pop1 > pop2 or vise versa



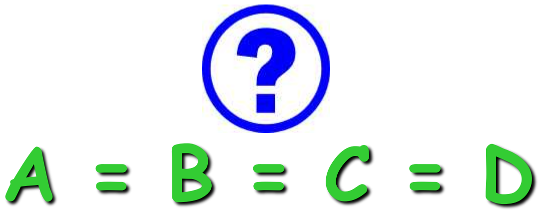
Basic Inferential Tests of Significance

How do you test for significance?

- t-Test - are two means different?



- paired t-Test - are means of paired data different?
  - Before after
- ANOVA - are any of a group of means different from the others?



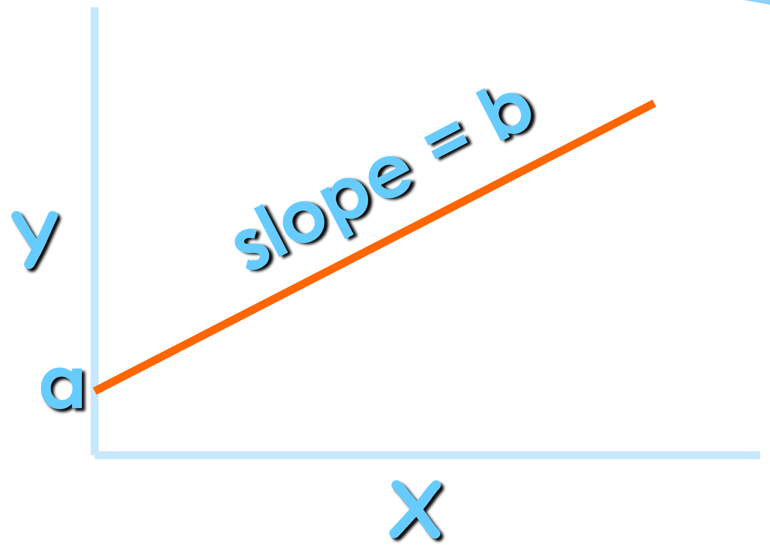
- Chi-square test - does observed freq. dist. differ from expected freq. dist.? X-test

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.

- $P=0.05$ – roughly 95% confident that your are not wrong
  - This has been determined to be the acceptable level of being wrong in most Science

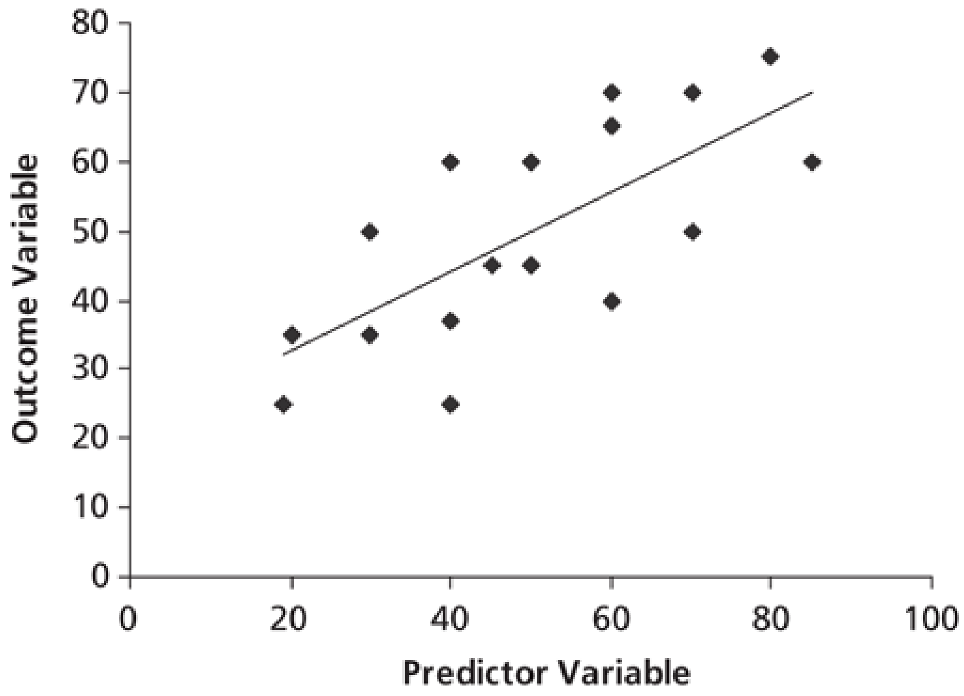
Regression Analysis and Measures of Association

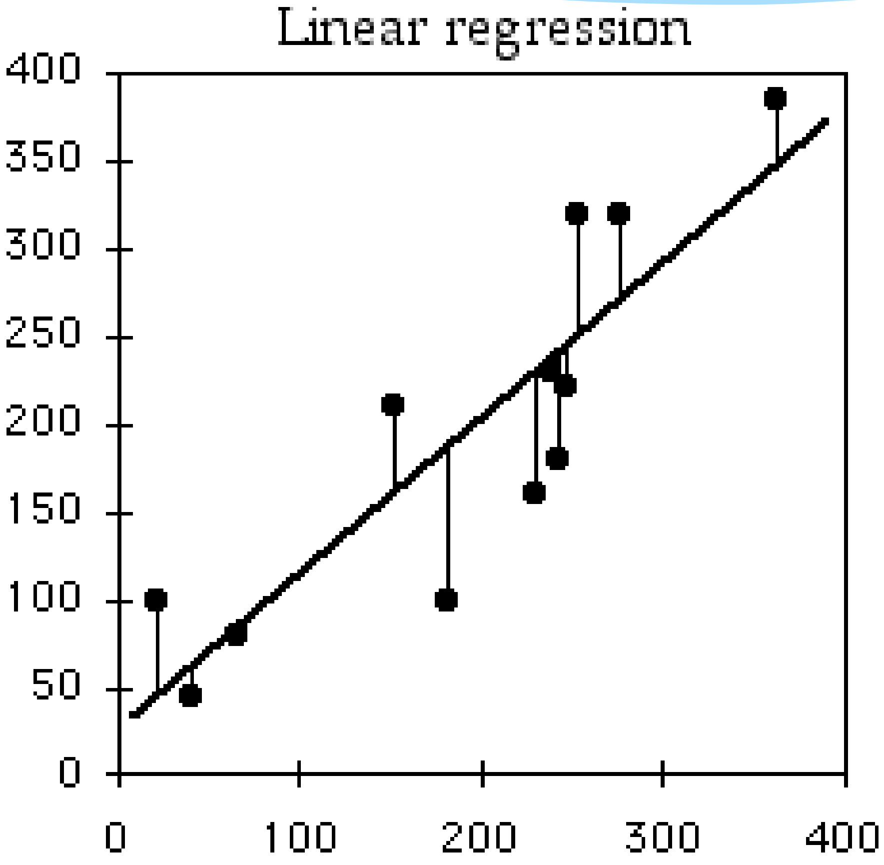


- linear regression - are two variables related according to y = a + b x
- correlation coefficient - ranges from -1 completely opposite to +1 completely similar

- Simple linear regression

Regression Analysis





Data transformations

- $\log_{10}$
- log e
- square
- square root
- sin
- cube

log (x)

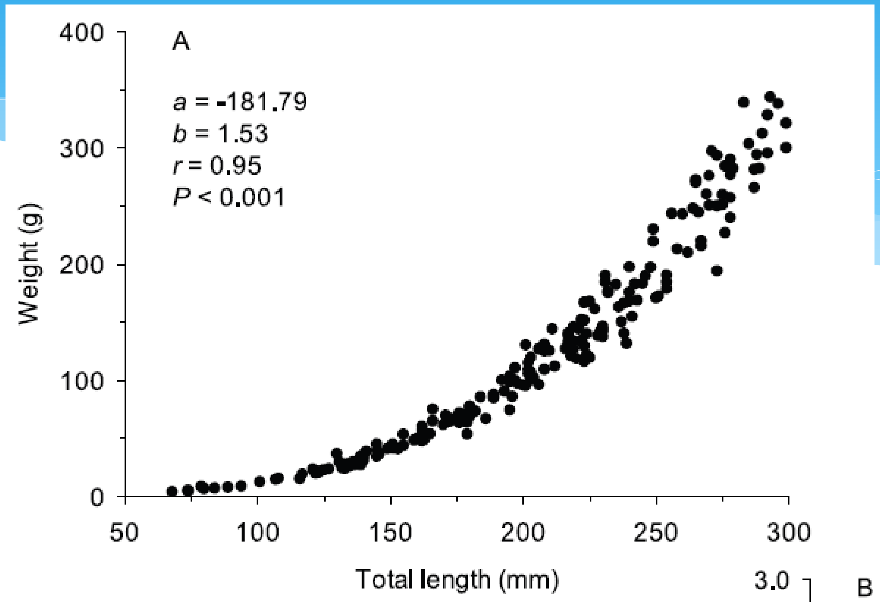
ln (x)

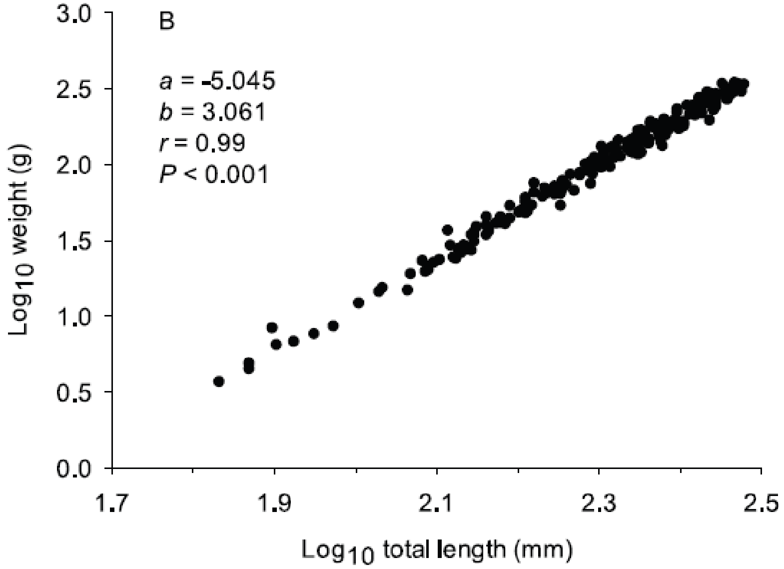
x^2

$\sqrt{x}$

sin (x)

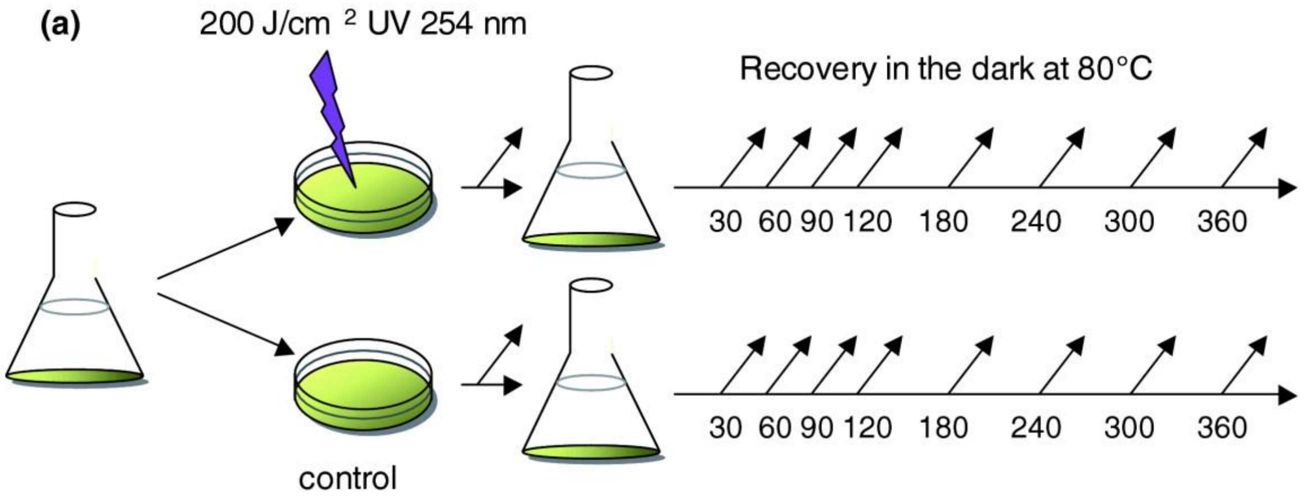
x^3





Critical Considerations in Study Design

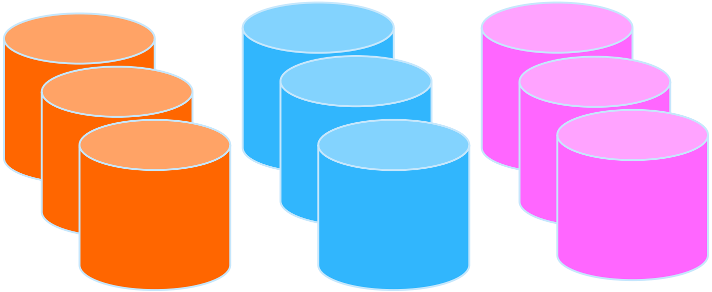
- Observational - passive monitoring over time or through space
- Experimental design – manipulate one variable
  - More than one treatment
  - one treatment is control



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



Self Check 7

- When trying to determine if two variables are correlated we could use Regression Analysis or simple linear regression.
  - **True**
  - False

- Passive monitoring over time or through space refers to which kind of experimental design
  - **Observational**
  - Experimental

Recap

- Data collection in the field
- Computer management
  - Electronic Data Collection & Databases
- Overview of stats
  - Descriptive
    - Central tendencies
    - Measures of dispersion
- Graphing data
  - Visualization is key
- Interpretation of data with statistics
  - Associations & Hypothesis testing