Population: the entire collection of individuals/objects of interest

Sample: a subset of the population

Types of Studies

- Experiment: researcher manipulates 1 or more independent variables (Factors)

  Treatments are the levels for the factors (or combinations thereof)

  Experimental Unit (EU)
  - the unit to which a treatment is randomly assigned
  - An EU constitutes one replication of the experiment

  Measurement Unit
  - the unit on which a measurement is taken
  - An EU constitutes one replication of the experiment

- Quasi-Experiment: treatments not randomly assigned to EUs

- Observational: no variables manipulated by researcher

  experiment may not be feasible or ethical

- Survey: voluntary response

Causal inference can be made from an experiment

Association relationships can be inferred from an observational study

Experimental Error - Variation among identically treated EUs

- Natural variation among EUs
- Measurement variability
- Variation in treatment conditions
- Extraneous factors (nuisance/lurking variables)
- Interaction of treatments and EUs

Control Treatments – A benchmark for comparing experimental treatments

- No treatment
- Placebo
- Standard practice

3 Principles of Designed Experiments

1) Blocking to reduce experimental error
2) Randomization to reduce hidden bias
3) Replication on an large number of subjects

Key Ingredients to identify:

A hypothesis
Dependent variable(s)
Experimental conditions
Nuisance variables
# of EUs
Assignment mechanism

Blocking - Grouping of EUs into similar classes

- Common Criteria for blocking
  - Location
  - Characteristics (age, weight, sex, …)
  - Time

Randomization - Random assignment of treatments to EUs

- Independent observations needed for valid estimates of experimental error
- Randomization simulates the effect of independence
  - Allows the assumption of independence & normal distribution

Replication

- Demonstrates reproducibility
- Allows for increased precision in estimating treatment effects

Surveys

- Administered to a sample from the population to gather information about the entire pop.
- Possible Problems:
  - non-response, incomplete recall, leading questions, unclear questions
  → Bias – when a study systematically favors certain outcomes

Sampling Designs for Surveys

- Simple Random Sample (SRS):

  A method of slecting *n* individuals from a pop. so that each is equally likely to be selected

- Stratified Random Sample:

  - Divide the population into groups of similar individuals (strata)

  - Take a Simple Random Sample from within each stratum

- Cluster Sampling:

  - Divide the population into groups of similar individuals (clusters)

  - Select a subset of clusters, and sample all individuals in the selected clusters

- Systematic Sampling:

  - Select every k^{th} individual from the population

  - May be more convenient, but is less efficient than other methods

  - Potential for bias is higher than for other methods

The National Health Interview Survey (NHIS)

- Conducted by the Census Bureau for the National Center for Health Statistics (NCHS)
  Uses:
  - to help set public policy
  - to track progress of national health objectives
  - to aid in research (in conjunction with the Medical Expenditure Panel Survey)

- Target Population: U.S. resident, civilian, non-institutionalized persons

- Sampling Frame: geographic areas defined in 3 stages

- Sampling Design: stratified multi-stage probability sample

- Components: Core Survey and usually 4 Supplements

1993 Sample:

- 43,007 households interviewed → 109,671 persons

- Non-interview rate: 4.4%

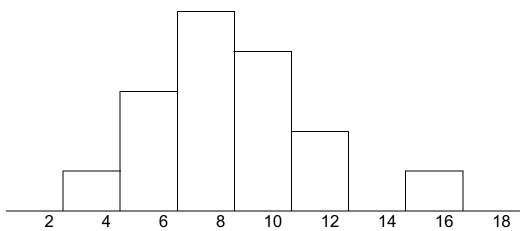
Histogram

1. Divide the range of values into equal length intervals
2. Count the # of observations in each interval
3. Plot adjacent rectangles with height equal to the count (or %) in each interval

- Example: Enzyme Data (concentrations sorted from lowest to highest)

2.3, 4.3, 4.7, 5.1, 6.3, 6.7, 6.9, 7.4, 7.8, 8.2, 8.2, 8.7, 9.6, 10.6, 10.8, 15.5

Interval	Count
2.1 – 4.0	1
4.1 – 6.0	3
6.1 – 8.0	5
8.1 – 10.0	4
10.1 – 12.0	2
12.1 – 14.0	0
14.1 – 16.0	1



Stem & Leaf Plot

- A histogram-like plot that allows you to recover the actual data

- The stem consists of all but the right most digit(s)
- The leaf consists of the last digit(s)

Procedure:
1. Write the stems in a column in increasing order
2. Use a vertical line to represent the decimal point
3. Write the leaves in increasing order next to the corresponding stem

Stem & Leaf Plot

- The stem consists of all but the right most digit(s)
- The leaf consists of the last digit(s)

stem	leaf
2	3
3	
4	37
5	1
6	379
7	48
8	227
9	6
10	68
11	
12	
13	
14	
15	5

(The decimal point is at the |)

Stem	Leaf	Count
15	5	1
14		
13		
12		
11		
10	68	2
9	6	1
8	227	3
7	48	2
6	379	3
5	1	1
4	37	2
3		
2	3	1

2|3 represents 2.3

Measures of Center

- Mean: ordinary average

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_i x_i}{n}$$

- Median (***M***): the "middle value" of the ordered data
  - ½ of the values are larger, ½ of the values are smaller
  - Procedure: Order the sample from smallest to largest
    - If n is odd, M is the middle value
    - If n is even, M is the average of the two middle values
  - The location of M among the ranked values is $(n+1)/2$

Example Data: 5, 4, 2, 6, 3

Measures of Spread

- Range: the difference between the largest and smallest values

- Quartiles: break a distribution into 4 intervals
  - ¼ of the observations are less than the 1st Quartile (Q1 = 25th Percentile)
  - ½ of the observations are less than the 2nd Quartile (Q2 = 50th Percentile)
  - ¾ of the observations are less than the 3rd Quartile (Q3 = 75th Percentile)

  Inter Quartile Range (IQR): the difference between the 3rd and 1st Quartiles
  - IQR = Q3 - Q1
  - Measures the spread of the middle half of the data

- 5 Number Summary: Minimum, Q1, Median, Q3, Maximum

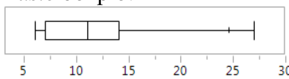
- Variance and Standard Deviation:

Ex Data: 7 7 8 9 11 12 12 14 15 16 28

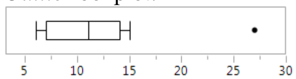
Example Data: 7 7 8 9 11 12 12 14 15 16 28

Boxplot: A graphical representation of a 5 number summary (developed by John Tukey)
- A central box at Q1 and Q3 with a line at the Median
- 'wiskers' extending to …
  - the Min and Max (*Basic* or *"Skeletal"* boxplot), **OR**
  - the lower and upper adjacent values (*Outlier* boxplot) [outliers indicated by *]

Basic boxplot



Outlier boxplot:



- Lower Inner Fence: $LIF = Q_1 - 1.5 * IQR$
- Upper Inner Fence: $UIF = Q_3 + 1.5 * IQR$

- **Outlier**: any observed value $< LIF$ or $> UIF$

- Lower Adjacent Value (*LAV*): the smallest non-outlier
- Upper Adjacent Value (*UAV*): the largest non-outlier