

- **Three ways to classify data**
 - **Quantitative vs. Qualitative**
 - * **Quantitative Data:** data that represents counts or measurements, answers the questions “how much?” or “how many?”, usually numerical
 - * **Qualitative Data:** data that separates units into categories by some non-numeric characteristic
 - **Discrete vs. Continuous**
 - * **Discrete Data:** any type of data where the possible values can be listed out completely
 - * **Continuous Data:** data where the possible values fall along an interval, and any list would miss many possible values
- **Types of Studies**
 - **Census:** A study of all of the units in the population
 - **Simulation:** Using a computer and thorough knowledge of the situation to model a setting and gather data about the setting without any “real world” work
 - **Observational Study:** A study where results are simply observed and measured

- **Experiment:** A study where the researcher has control over subjects and assigns some treatment before observing the effect
 - * **Completely Randomized Design (CRD):** each subject is assigned to a treatment group completely at random without regard to any other factors
 - * **Confounding:** this is what happens if the effect of two factors cannot be distinguished from each other
 - * **Blocking:** putting subjects randomly into different groups according to some important factor, so that each level of the treatment is represented fairly across the levels of the blocking factor
 - * **Placebo:** a way of studying the psychological effects of a treatment, looks, tastes, feels, etc. just like the real treatment
 - * **Blinding:** not letting the subject know whether they are in a treatment or placebo group, in medical studies the patient's doctor also is not informed of which group the patient is in so that they cannot hint at the truth (this is called a double blind experiment)
- **A Biased Sample:** a sample in which the units selected or the data gathered are not representative of the population as a whole

- **Sampling Methods**

- **Random Sample:** each subject has the same chance of being selected
- **Simple Random Sample (SRS):** each sample of size n is equally likely
- **Systematic Sampling:** take every k^{th} subject after some random starting point
- **Stratified Sampling:** subdivide the population into several large subgroups (or strata) and conduct a SRS within each
- **Cluster Sampling:** subdivide the population into many small clusters, then randomly select some of the clusters and use all of the subjects within each selected cluster
- **Convenience Sampling:** use whichever subjects are easiest to get

- **Two Types of Errors Involved with Collecting Data**

- **Sampling Error:** the difference between a sample result and the true population that is due entirely to random variations
- **Nonsampling Error:** refers to all variations that are not due to random chance such as non-random samples, misrecorded data, subjects lying on surveys, poorly worded questions, etc.

- **Probability Experiment:** an action through which results are obtained with various possible results

- **Outcome:** one of the possible results
- **Sample Space:** all possible outcomes
- **Event:** a collection of one or more possible outcomes
- **Probability:** we will think of probabilities as closely related to the idea of relative frequency, that is a probability expresses the long term proportion of the time some event would occur if the probability experiment were run over and over again (even if this is not actually possible)
- **Law of Large Numbers:** as our sample size gets larger, our sample relative frequencies will become better and better approximations of the true probabilities
- **Limits:** all probabilities must be between 0 and 1
- **Interpreting:** a probability of 0 means an event is impossible, while 1 means it is certain
- **Complement:** the complement of an event is the set of all outcomes not in the original event, thus the probability of an event and its complement must add up to 1
- **Independence:** two events are independent if knowledge of whether one occurs does not affect the probability that the other occurs, if this is the case, the probability that both events occur is found by multiplying the probabilities of each individual event
- **Mutually Exclusive:** two events are called mutually exclusive (or disjoint) if they cannot occur together

- **Random Variable:** a number whose value is determined by some random process whose behavior can be modeled using probabilities
- **Types of Distributions:** random variables are generally broken down into discrete and continuous types
- **Features of Discrete Distribution:** a function called a probability mass function (pmf) assigns probability to each of the possible values such that the assigned probabilities add up to 1 when summed over all possible values
- **Common examples:** common discrete distributions include the geometric, the hypergeometric, the Poisson, the Bernoulli, and the binomial, each of which is actually a collection of related distributions called a “family”; we will only comment on the binomial family here
- **Traits of a “Binomial Distribution”:**
 - There are a fixed number (n) of trials
 - The trials are independent of each other
 - There are two possible outcomes for each trial
 - The random variable is the count of one of these outcomes (we’ll call the outcome we’re counting a “success”)
 - The probability of success is the same from one trial to the next trial