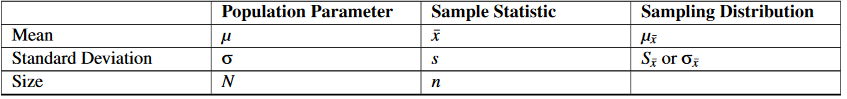
* **Sample** = a *representative* subset of a **population**.
* To get info about a population, the only way to be truly sure is to conduct a **census,** which is impractical in most situations 🡪 extremely expensive, labor-intensive, + time-consuming to organize, interpret, + display the info.
* Also, by the time the results are published, it would be very probable events could change the population + the results would be obsolete.
* In addition, a census has the potential to be destructive to the population being studied.
* Ex: manufacturing company testing padlocks for quality control via a machine to see how much force it can apply to the lock before it breaks.
* If they did this with every lock, they would have none left to sell
* Ex: a biologist finds the # of fish in a lake by draining the lake + counting them all
* Due to all of the difficulties associated w/ a census, **sampling** is much more practical
* But even the most carefully planned sample will be subject to random variation between the sample + the population
* These differences due to chance = **sampling error**, or the **margin of error**
* Also, it’s often difficult to obtain a sample that *accurately* reflects the total population.
* It is also possible to make mistakes in selecting the sample and/or in collecting the info.
* These problems result in a non-representative sample, one in which our conclusions differ from what they would have been if we had been able to conduct a census.
* The term most frequently applied to a non-representative sample is **bias,** whichhas many potential sources.
* It is important when selecting a sample/designing a survey that a statistician make every effort to eliminate potential sources of bias.
* **Sampling Bias 🡪** the methods used in selecting the sample.
* The **sampling frame =** the group/listing from which the sample is to be chosen.
* If you wanted to study the population of students in your school, you could obtain a list of all the students from the office + choose students from the list = the sampling frame.
* **Incorrect Sampling Frame**🡪 If the list from which you choose your sample does NOT accurately reflect the characteristics of the population
* A sampling frame error occurs when some group from the population does not have the opportunity to be represented in the sample.
* Ex: surveys over telephone 🡪 use telephone book as a sampling frame by choosing #’s from it, but some #’s are not listed in the telephone book.
* In addition, younger adults in particular tend to only use cell phones or CPU-based phone services + may not even have traditional phone service.
* Even if you picked #’s randomly, the sampling frame could be incorrect, b/c there are also people who have no phone 🡪 absolutely no chance to be represented in the sample.
* **Undercoverage 🡪** describes the problems when a group of the population is not represented in a survey + can result from all of the different sampling biases.
* Ex: 1936 presidential election. Literary Digest obtained a huge sample of 10M people, + from that pool, 2M replied + they predicted Landon would win the election.
* As it turned out, the election was won in a landslide by FDR
* W/ these #’s, you’d typically expect very accurate results, but the magazine used their own subscriber list as their sampling frame.
* During the depression, these individuals would have been only the wealthiest Americans, who tended to vote Republican, and left the majority of typical voters under-covered.
* **Convenience Sampling**
* Can cause certain opinions to be over-represented in relation to the true population or undercoverage.
* **Judgment Sampling**
* Occurs when an individual/organization usually considered an expert in the field being studied chooses the individuals or group of individuals to be used in the sample.
* B/c it is based on a *subjective* choice, even by someone considered an expert, it is very susceptible to bias.
* In some sense, this is what those responsible for the Literary Digest poll did 🡪 incorrectly chose groups they believed would represent the population.
* If a person wants to do a survey on the middle-class, how would they decide who to include?
* It would be left to this person’s *own judgment* to create criteria for “middle-class”, which might differ + include wealthier individuals that others would not consider part of the population.
* **In Quota Sampling 🡪** an individual/organization attempts to include the proper proportions of individuals of different subgroups in their sample.
* Might sound like a good idea, but is subject to individual’s prejudice + is therefore prone to bias.
* **Size Bias**
* If 1 particular subgroup in a population is likely to be over/under-represented due to its size
* If we chose a state at random from a map by closing our eyes + pointing to a particular place, larger states have a greater chance of being chosen than smaller ones.
* Ex: survey to find out typical size of a student’s math class at a school 🡪 chances are greater that we’d choose someone from a larger class for our survey.
* **Response Bias =** problems resulting from ways in which a survey/poll is presented to the sample.
* **Voluntary Response Bias**
* TV + radio stations asking viewers/listeners to call in w/ opinions about a particular issue, online poll question of the day, Reality TV + fan balloting in all-star sports players
* These types of polls are very susceptible to voluntary/**self-selection bias**
* People who respond to these types of surveys tend to feel very strongly 1 way or another about the issue in question, + results might not reflect the overall population.
* Those who still have an opinion, but may not feel *quite* so passionately about the issue + may not be motivated to respond to the poll.
* **Non-Response Bias**
* 1 of the biggest problems in polling = most people don’t want to be bothered to respond
* Don’t know how much these individuals’ beliefs + opinions reflect those of a general population, +, therefore, almost all surveys could be prone to non-response bias
* **Questionnaire Bias**
* occurs when the way in which a question is asked influences responses given
* It’s possible to ask the same question in 2 different ways that would lead individuals w/ the same basic opinions to respond differently.
* Consider the following 2 questions about gun control.
* "Do you believe it is reasonable for the government to impose some limits on purchases of certain types of weapons in an effort to reduce gun violence in urban areas?"
* "Do you believe it is reasonable for the government to infringe on an individual’s constitutional right to bear arms?"
* A gun rights activist might feel very strongly that the government should never be in the position of limiting guns in any way + would answer NO to both questions.
* Someone who is very strongly against gun ownership, on the other hand, would probably answer YES to both questions.
* However, individuals w/ a more tempered, middle position might believe in an individual’s right to own a gun under some circumstances, while still feeling a need for regulation.
* These individuals would most likely answer these two questions differently.
* Questionnaire bias is not necessarily always a deliberate action.
* If a question is poorly worded, confusing, or just plain hard to understand, it could lead to non-representative results.
* When you ask people to choose between 2 options, it is even possible *the order* in which you list the choices may influence their response!
* **Incorrect Response Bias** 🡪 an individual intentionally responds to a survey w/ an untruthful answer
* Major problem w/ surveys = you can never be sure the person is actually responding truthfully.
* Can occur when asking questions about extremely sensitive or personal issues, like illegal drinking among teens
* Even if guaranteed that responses are confidential, some may not want to admit to engaging in certain behaviors at all + or will lie to appear different than they really are
* Identifying Sources of Bias
* Example A: Study attempting to determine satisfaction of school communication w/ students who speak a 2nd language at home
* Plan: send home a questionnaire to the parents of the students asking about their opinion.
* This is liable to result in both **non-response** + **undercoverage** bias.
* Any time a sample is expected to submit a questionnaire, results are going to include more input from the type of person willing + able to complete + submit the survey.
* Undercoverage is a particular problem b/c the population most affected by the study is also unusually liable to misinterpret questions/the reason for them due to the language barrier.
* Possible solution: conduct phone survey by a native speaker in the target language(s)
* Example B
* An experiment to determine the danger of mixing household chemicals is conducted by collecting samples of chemicals found under the experimenter’s sink
* **Undercoverage due to convenience** **🡪** chemicals used were the ones conveniently found in 1 location so results could not be assumed to be the same as w/ chemicals found under other sinks.
* Mall shoppers are asked to fill out and return a form rating their shopping experiences at each of the 26 stores to ID the most popular stores in each of 4 categories
* **Non-response bias 🡪** not all shoppers will respond so results will be biased toward a specific type of personality + won’t reflect a true cross-section of shoppers’ experiences.
* Study of the average grades of mathematics students polls 16 Algebra I students, 14 Geometry students, 7 Calculus students, and 19 Statistics students.
* ½ as many Calculus students as the other subjects = **undercoverage**
* The best technique for reducing bias in sampling is **randomization**
* When a **simple random sample** of size n is taken from a population, all possible samples of size n in the population have an equal probability of being selected for the sample.
* It is important to note *there is no such thing as true randomness*, especially on a calculator or CPU
* When you choose the ’rand’ function, the calculator has been programmed to return a ten digit decimal that, using a very complicated mathematical formula, simulates randomness.
* Each digit, in theory, is equally likely to occur in any of the individual decimal places.
* What this means in practice is if you had patience (+ time) to generate a million of these + keep track of the frequencies in a table, you’d find an approximately equal # of each digit.
* However, 2 brand-new calculators will give the exact same sequences of random numbers b/c the function that simulates randomness has to start at some number, called a **seed value**
* All calculators are programmed (or when memory is reset) to use a seed value = 0
* If you want to be sure your sequence of random digits is different from everyone else’s, seed your random number function using a number different from theirs.
* **Systematic Sampling** 🡪 after choosing a starting point at random, subjects are selected using a **jump number**
* Jump is determined by dividing population size by the desired sample size to insure the sample combs through the entire population.
* Ex: choosing teams/groups in gym class by counting off by 5 via generating a random # from 1 to 25 (n = 25) 🡪 get 14 as our seed value 🡪 start w/ student 14 + select every 5th student until we had 5 in all.
* When we came to the end of the list, continue the count at number 1.
* Thus, our chosen students would be: 14, 19, 24, 4, 9.
* It is important to note this is NOT a simple random sample, as not *every* possible sample of 5 students has an equal chance of being chosen.
* For example, it is impossible to have a sample consisting of students 5, 6, 7, 8, and 9
* **Cluster sampling 🡪** a naturally occurring group is selected at random, + then either all of that group or randomly selected individuals from that group are used for the sample.
* **multi-stage sampling =** select at random from out of that group, or cluster into smaller subgroups
* Ex: survey student opinions 🡪 choose 5 schools at random from a state + use SRS in each
* If we wanted a national survey of urban schools 🡪 choose 5 major urban areas from around the country at random 🡪 select 5 schools at random from each of these cities.
* This would be BOTH cluster + multi-stage sampling.
* Cluster sampling is often done by selecting a particular block/street at random from w/in a town/city, or at large public gatherings or rallies.
* Officials take a small, representative area of the crowd + count the individuals in just that area, they can use that count to estimate the total crowd in attendance.
* **Stratified Sampling** 🡪 population is divided into groups, called **strata**, that have some meaningful relationship
* Very often, groups in a population that are similar may respond differently to a survey.
* In order to help reflect the population, stratify to insure each opinion is represented in a sample.
* Often stratify by gender or race in order to make sure the often divergent views of these different groups are represented.
* In a survey of high school students, might stratify by school to be sure opinions of different communities are included.
* If each school has an approximately equal number of students, could choose to take an SRS of size 25 from each school.
* If the numbers in each stratum are different, it would be more appropriate to choose a fixed sample (100 students, for example) from each school + take a number from each school proportionate to the total school size
* Statistics allows us to make use of probability to estimate what is true from just a sample of the subjects we are interested in.
* Increasing sample size results in a sampling distribution of means more closely clustered around the true mean, is normal-looking, + improves estimates
* As we increase sample size, sample means become more + more normally distributed around the true mean (the population parameter) + variability of the sample means/**standard error** decreases (more tightly clustered around the true mean.
* **SE** = SD of a sampling distribution 🡪 **Sigma / Sqrt(n)**
* The sampling distribution, as it becomes more normal in shape, also adheres to the **Empirical Rule**.
* This means certain proportions of the sample means will fall w/in defined increments, where each increment = 1 SE from the population parameter.
* According to this rule, 34% of the sample means will fall within 1 SE standard error + below the population parameter
* Probability theory says 95% of the samples will fall within 2 SE of the true value, 99.7% fall w/in 3
* As sample size increases, the shape of the sample means distribution becomes more + more Normal, so although the true mean of the population can be unknown, random sampling can yield a reliable estimate of it.
* **Central Limit Theorem 🡪** as you increase sample size for a random variable, the distribution of the sample means better approximates a normal distribution.
* This theorem is important b/c it allows us to:
* predict true population mean w/in a defined degree of **confidence** if we can select a single sample of a known size from our population + calculate its mean
* *This holds true no matter what the shape of the original distribution.*



* Formally, the CLT says:
* *If samples of size n are drawn at random from any population w/ a finite mean + SD, the sampling distribution of the sample means, x(bar), approximates a normal distribution as n increases*
* So if you collect many samples from an ordinary random variable + calculate the mean of each sample, the means will be distributed in an approximate bell-curve + the “**mean of means**” will be the same as the mean of the population.
* The larger the size of the samples you collect, the more closely the distribution of their means will approximate a normal distribution.
* How can we use the CLT to help us construct a sampling distribution w/out repeatedly sampling?
* Use what we know about the population + our proposed sample size to sketch the theoretical sampling distribution.
* Remember, to sketch a distribution we need to know its shape, center and spread.
* **Notes to remember:**
* As long as sample size is at least 30, you may assume the distribution of the sample means to be approximately normal (true regardless of the original distribution of the random variable)
* The mean of a sampling distribution = mean of the population (**mu = mu(x)**)
* The SD of the sample means = the SE + can be estimated by dividing SD of the population by the square root of the sample size (**S(x) = sigma / Sqrt(n)**)
* Ex: The time it takes a student to complete the mid-term for Algebra II is a bi-modal distribution with mu = 1 hour + sigma = 1 hours. During the month of June, Professor Spence administers the test 64 times. What is the probability that the average mid-term completion time for students during the month of June exceeds 48 minutes?
* Important facts:
* There are more than 30 samples, so the CLT applies.
* The mean of the sample should approximate the mean of the population 🡪 mu(x) = mu
* The SD of Professor Spence’s sample = 1 / sqrt(64) = 1/8
* 48 minutes = 48/60 = 0.8 hours 🡪 the range we are interested in is x > 0.8
* Calculate z-score of minimum value in the relevant range, 0.80 hours **🡪 (0.8 – 1) / 0.125 = -1.6**
* P(Z >= -1.6) from z-table = .9452 = 94.52% chance the average mid-term completion time for students during the month of June exceeds 48 minutes
* Ex: Evan price-checked 123 online sellers to record average asking price for his favorite game. According to a major nation price-checking site, national average online cost for the game is $35 w/ SD = $3 a standard. Evan found the prices less than $34.86 on average. How likely is this result?
* Since there are more than 30 samples, apply the CLT + treat the sample as a normal distribution.
* mu(x) should = mu SE = 3 / sqrt(123) = 0.27
* z-score of Evan’s found price averaged 🡪 (34.86 – 35) / 0.27 = -0.518 = -0.52
* Check z-table 🡪 .3015 🡪 30.15% chance of this result
* Ex: Mack asked 42 fellow students how much they spent for lunch, on average. According to his research, the amount spent by students nationwide = $15 w/ SD = $9. We assume Mark’s random sample should fall w/in this sampling distribution. What is the probability that Mack’s random sample will have a value w/in $0.01 of the national average?
* Want probability Mack’s sample has a value between $14.99 and $15.01
* 42 ≥ 30 🡪 CLT 🡪 can safely assume the sampling distribution of the sample mean will be approximately normal
* mu(x) should = mu SE = 9 / sqrt(42) = 1.389
* z-score of lower bound 🡪 (14.99 – 15) / 1.389 = -0.0072 🡪 -.01
* z-score of upper bound 🡪 (15.01 – 15) / 1.389 = 0.0072 🡪 .01
* z-table 🡪 .4602 and .5398 🡺 .5398 - .4602 = 0.0796 = 7.96% chance of being in this range