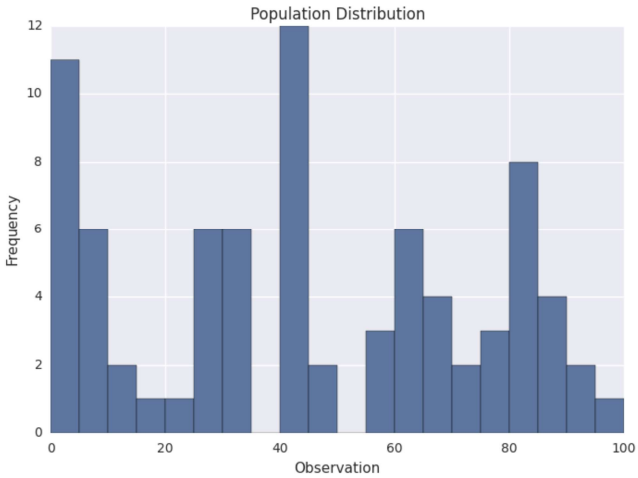
Statistics

Q1. What is the Central Limit Theorem and why is it important?

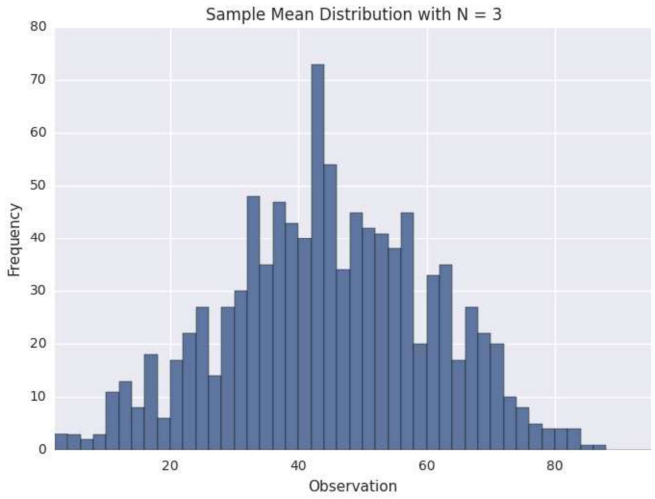
https://spin.atomicobject.com/2015/02/12/central-limit-theorem-intro/

Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impractical, bordering on impossible. While we can't obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample.
The Central Limit Theorem addresses this question exactly. Formally, it states that if we sample from a population using a sufficiently large sample size, the mean of the samples (also known as the sample population) will be normally distributed (assuming true random sampling), the mean tending to the mean of the population and variance equal to the variance of the population divided by the size of the sampling. What's especially important is that this will be true regardless of the distribution of the original population.

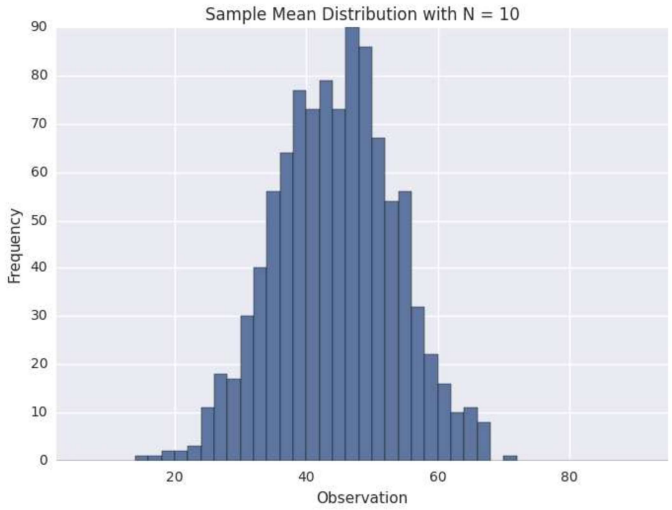
EX:



As we can see, the distribution is pretty ugly. It certainly isn't normal, uniform, or any other commonly known distribution. In order to sample from the above distribution, we need to define a sample size, referred to as N. This is the number of observations that we will sample at a time. Suppose that we choose N to be 3. This means that we will sample in groups of 3. So for the above population, we might sample groups such as [5, 20, 41], [60, 17, 82], [8, 13, 61], and so on.
Suppose that we gather 1,000 samples of 3 from the above population. For each sample, we can compute its average. If we do that, we will have 1,000 averages. This set of 1,000 averages is called a sampling distribution, and according to Central Limit Theorem, the sampling distribution will approach a normal distribution as the sample size N used to produce it increases. Here is what our sample distribution looks like for N = 3.



As we can see, it certainly looks uni-modal, though not necessarily normal. If we repeat the same process with a larger sample size, we should see the sampling distribution start to become more normal. Let's repeat the same process again with N = 10. Here is the sampling distribution for that sample size.



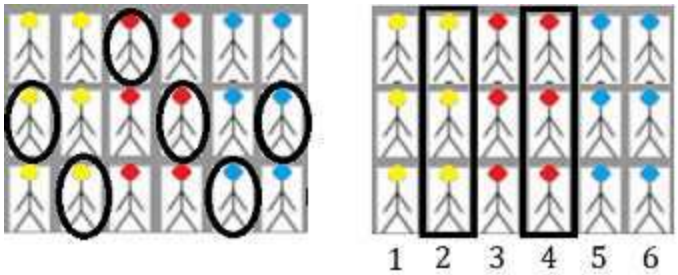
Q2. What is sampling? How many sampling methods do you know?

https://searchbusinessanalytics.techtarget.com/definition/data-sampling
https://nikolanews.com/difference-between-stratified-sampling-cluster-sampling-and-quota-sampling/

Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined. It enables data scientists, predictive modelers and other data analysts to work with a small, manageable amount of data about a statistical population to build and run analytical models more quickly, while still producing accurate findings.

Sampling can be particularly useful with data sets that are too large to efficiently analyze in full – for example, in big data analytics applications or surveys. Identifying and analyzing a representative sample is more efficient and cost-effective than surveying the entirety of the data or population.
An important consideration, though, is the size of the required data sample and the possibility of introducing a sampling error. In some cases, a small sample can reveal the most important information about a data set. In others, using a larger sample can increase the likelihood of accurately representing the data as a whole, even though the increased size of the sample may impede ease of manipulation and interpretation.
There are many different methods for drawing samples from data; the ideal one depends on the data set and situation. Sampling can be based on probability, an approach that uses random numbers that correspond to points in the data set to ensure that there is no correlation between points chosen for the sample. Further variations in probability sampling include:

- Simple random sampling: Software is used to randomly select subjects from the whole population.
- Stratified sampling: Subsets of the data sets or population are created based on a common factor, and samples are randomly collected from each subgroup. A sample is drawn from each strata (using a random sampling method like simple random sampling or systematic sampling).
    - *EX: In the image below, let's say you need a sample size of 6. Two members from each group (yellow, red, and blue) are selected randomly. Make sure to sample proportionally: In this simple example, 1/3 of each group (2/6 yellow, 2/6 red and 2/6 blue) has been sampled. If you have one group that's a different size, make sure to adjust your proportions. For example, if you had 9 yellow, 3 red and 3 blue, a 5-item sample would consist of 3/9 yellow (i.e. one third), 1/3 red and 1/3 blue.*
- Cluster sampling: The larger data set is divided into subsets (clusters) based on a defined factor, then a random sampling of clusters is analyzed. The sampling unit is the whole cluster; Instead of sampling individuals from within each group, a researcher will study whole clusters.
    - *EX: In the image below, the strata are natural groupings by head color (yellow, red, blue). A sample size of 6 is needed, so two of the complete strata are selected randomly (in this example, groups 2 and 4 are chosen).*

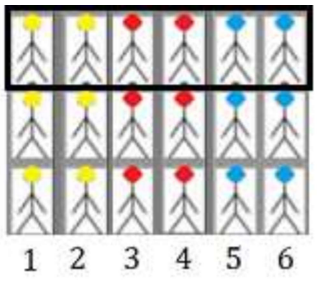


- Multistage sampling: A more complicated form of cluster sampling, this method also involves dividing the larger population into a number of clusters. Second-stage clusters are then broken out based on a secondary factor, and those clusters are then sampled and analyzed. This staging could continue as multiple subsets are identified, clustered and analyzed.
- Systematic sampling: A sample is created by setting an interval at which to extract data from the larger population – for example, selecting every 10th row in a spreadsheet of 200 items to create a sample size of 20 rows to analyze.

Sampling can also be based on non-probability, an approach in which a data sample is determined and extracted based on the judgment of the analyst. As inclusion is determined by the analyst, it can be more difficult to extrapolate whether the sample accurately represents the larger population than when probability sampling is used.

Non-probability data sampling methods include:

- Convenience sampling: Data is collected from an easily accessible and available group.
- Consecutive sampling: Data is collected from every subject that meets the criteria until the predetermined sample size is met.
- Purposive or judgmental sampling: The researcher selects the data to sample based on predefined criteria.
- Quota sampling: The researcher ensures equal representation within the sample for all subgroups in the data set or population (random sampling is not used).



Quota Sampling

Once generated, a sample can be used for predictive analytics. For example, a retail business might use data sampling to uncover patterns about customer behavior and predictive modeling to create more effective sales strategies.

Q3. What is the difference between type I vs type II error?

https://www.datasciencecentral.com/profiles/blogs/understanding-type-i-and-type-ii-errors

Is H_a true? No, H_0 is True (H_a is Negative: TN); Yes, H_0 is False (H_a is Positive: TP).
A type I error occurs when the null hypothesis is true but is rejected. A type II error occurs when the null hypothesis is false but erroneously fails to be rejected.

	No reject H_0	Reject H_0
H_0 is True	TN	FP (I error)
H_0 is False	FN (II error)	TP

Q4. What is linear regression? What do the terms p-value, coefficient, and r-squared value mean? What is the significance of each of these components?

https://www.springboard.com/blog/linear-regression-in-python-a-tutorial/
https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients

Imagine you want to predict the price of a house. That will depend on some factors, called independent variables, such as location, size, year of construction… if we assume there is a linear relationship between these variables and the price (our dependent variable), then our price is predicted by the following function:

$$Y = a + b\,X$$

The p-value in the table is the minimum α (the significance level) at which the coefficient is relevant. The lower the p-value, the more important is the variable in predicting the price. Usually we set a 5% level, so that we have a 95% confidentiality that our variable is relevant.
The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. This property of holding the other variables constant is crucial because it allows you to assess the effect of each variable in isolation from the others.

R squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Q5. What are the assumptions required for linear regression?

There are four major assumptions:

- There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data,
- The errors or residuals ($y_i - \hat{y}_i$) of the data are normally distributed and independent from each other,
- There is minimal multicollinearity between explanatory variables, and
- Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

Q6. What is a statistical interaction?

http://icbseverywhere.com/blog/mini-lessons-tutorials-and-support-pages/statistical-interactions/

Basically, an interaction is when the effect of one factor (input variable) on the dependent variable (output variable) differs among levels of another factor. When two or more independent variables are involved in a research design, there is more to consider than simply the "main effect" of each of the independent variables (also termed "factors"). That is, the effect of one independent variable on the dependent variable of interest may not be the same at all levels of the other independent variable. Another way to put this is that the effect of one independent variable may depend on the level of the other independent variable. In order to find an interaction, you must have a factorial design, in which the two (or more)

independent variables are "crossed" with one another so that there are observations at every combination of levels of the two independent variables. *EX: stress level and practice to memorize words: together they may have a lower performance.*

Q7. What is selection bias?

https://www.elderresearch.com/blog/selection-bias-in-analytics

Selection (or 'sampling') bias occurs when the sample data that is gathered and prepared for modeling has characteristics that are not representative of the true, future population of cases the model will see. That is, active selection bias occurs when a subset of the data is systematically (i.e., non-randomly) excluded from analysis.

Q8. What is an example of a data set with a non-Gaussian distribution?

https://www.quora.com/Most-machine-learning-datasets-are-in-Gaussian-distribution-Where-can-we-find-the-dataset-which-follows-Bernoulli-Poisson-gamma-beta-etc-distribution

The Gaussian distribution is part of the Exponential family of distributions, but there are a lot more of them, with the same sort of ease of use, in many cases, and if the person doing the machine learning has a solid grounding in statistics, they can be utilized where appropriate.

Binomial: multiple toss of a coin Bin(n,p): the binomial distribution consists of the probabilities of each of the possible numbers of successes on n trials for independent events that each have a probability of p of occurring.
Bernoulli: Bin(1,p) = Be(p)
Poisson: Pois(λ)