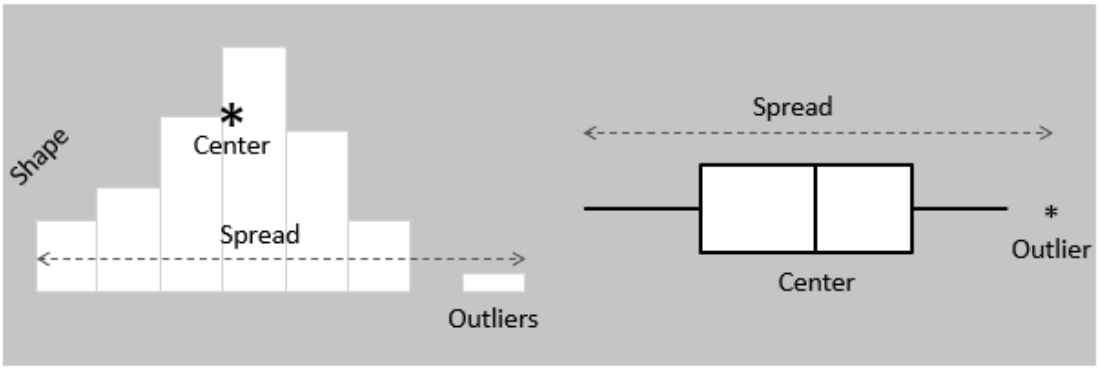
STATISTICS WORKSHEET-6

1. d) all of the mentioned
2. a) Discrete
3. a) pdf
4. c) mean
5. b),c)
6. a) variance
7. c) 0 and 1
8. b) bootstrap
9. b) summarized

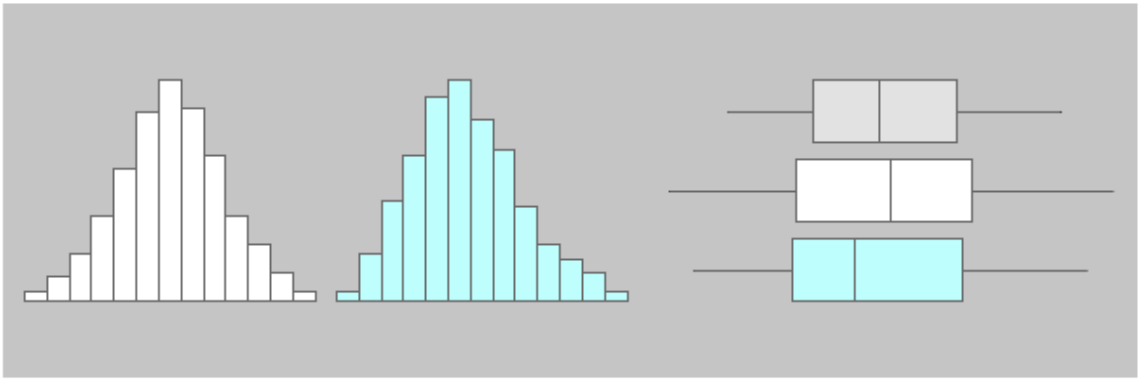
10. Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the inter quartile range of the data. The 'whiskers' of a box plot show the least and greatest values in the data set.

Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.

Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.



Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.



THE PROS AND CONS OF USING A HISTOGRAM VS A BOX PLOT

Histogram

Pros :

1. It divides the numeric data into uniform intervals and displays the number of data values falling within each bin.

2. They group data into a small chunk. They are useful for summarizing numeric data in that they show the rough distribution of values.

cons

1. The histogram doesn't show information about what is happening within each bin of the graph.
2. It shows the number of values within an interval but not the actual values.

Box Plot

Pros:

1. It is a good way to summarize large amounts of data.

2. It is easier to read minimum value, median, outliers, quantiles, and maximum value.

Cons

1. It shared to identify the original data we will use box plot if I have display the range and distribution of data whereas histogram will be used to displays the number of values within an interval.

11. Metrics are **measures of quantitative assessment commonly used for comparing, and tracking performance or production**. Metrics can be used in a variety of scenarios. Metrics are heavily relied on in the financial analysis of companies by both internal managers and external stake holders. Good metrics offer two things: 1) They help you identify how your business is doing and 2) They tell you what to focus on. That's why it's important to have them prominently displayed. They keep you on track!

12. Statistical significance refers to the likelihood that a relationship between two or more variables is not caused by random chance. In essence, it's a way of proving the reliability of a certain statistic. Its two main components are sample size and effect size. In the use of statistical hypothesis testing, a data set's result can be deemed statistically significant if you have reached a certain level of confidence in the result. In statistical hypothesis testing, this means the hypothesis is unlikely to have occurred given the null hypothesis. According to a null hypothesis, there is no relationship between the variables in question. To assess statistical significance, we would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, calculate the p-value, which is the likelihood of getting the test's observed findings if the null hypothesis is true. Finally, you would select the threshold of significance (alpha) and reject the null hypothesis if the p-value is smaller than the alpha — in other words, the result is statistically significant.

calculate statistical significance

1. Create a null hypothesis.
2. Create an alternative hypothesis.
3. Determine the significance level.
4. Decide on the type of test you'll use.
5. Perform a power analysis to find out your sample size.
6. Calculate the standard deviation.
7. Use the standard error formula.
8. Determine the t-score.
9. Find the degrees of freedom.
10. Use a t table.

13. Any type of categorical data won't have a gaussian distribution or log-normal distribution. The normal distribution takes center stage in statistics, many processes follow a **non normal distribution**. This can be due to the data naturally following a specific type of non normal distribution .**Exponential distributions** - eg.

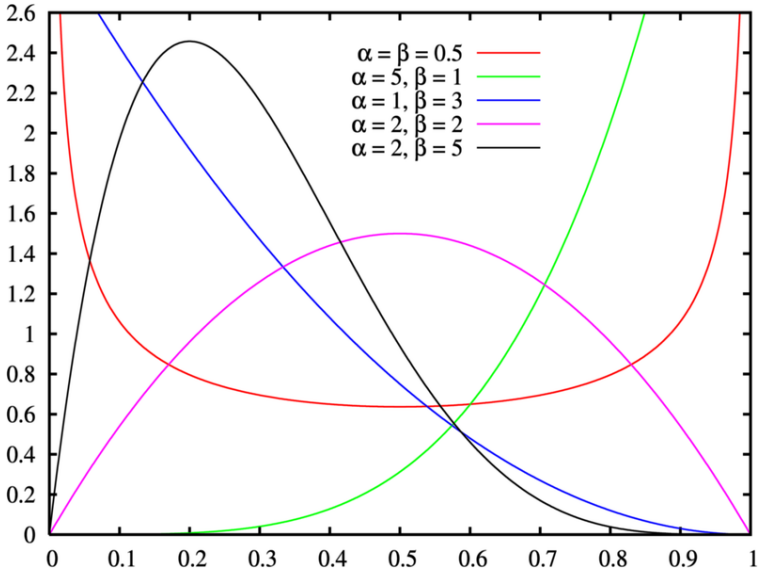
- The amount of time that a car battery lasts.

- The amount of time until an earthquake occurs.

- Allocation of wealth among individuals
- Values of oil reserves among oil fields (many small ones, a small number of large ones)

Types of Non Normal Distribution

Many distributions naturally follow non normal patterns.



1. Beta Distribution.
2. Exponential Distribution.
3. Gamma Distribution.
4. Inverse Gamma Distribution.
5. Log Normal Distribution.
6. Logistic Distribution.
7. Maxwell-Boltzmann Distribution.
8. Poisson Distribution.
9. Skewed Distribution.
10. Symmetric Distribution.
11. Uniform Distribution.
12. Unimodal Distribution.
13. Weibull Distribution.

14. The **mean** of a dataset represents the average value of the dataset. It is calculated as:

Mean = Σx_i / n

where:

- **Σ:** A symbol that means “sum”
- **x_i:** The i^{th} observation in a dataset
- **n:** The total number of observations in the dataset

The **median** represents the middle value of a dataset. It is calculated by arranging all of the observations in a dataset from smallest to largest and then identifying the middle value.

For example, suppose we have the following dataset with 11 observations:

Dataset: 3, 4, 4, 6, 7, 8, 12, 13, 15, 16, 17

The mean of the dataset is calculated as:

Mean = (3+4+4+6+7+8+12+13+15+16+17) / 11 = **9.54**

The median of the dataset is the value directly in the middle, which turns out to be **8:**

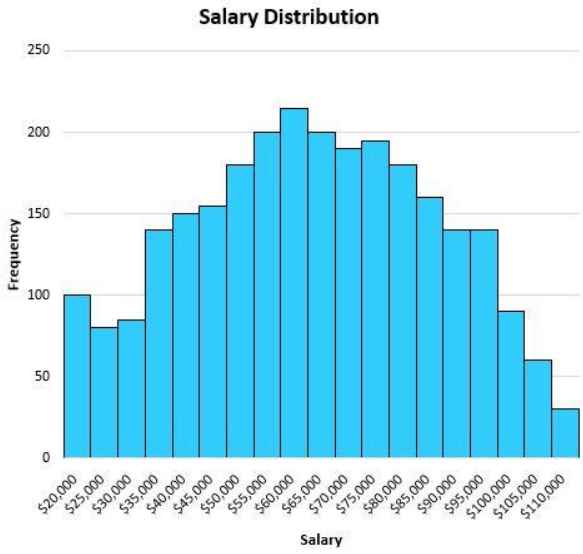
3, 4, 4, 6, 7, **8**, 12, 13, 15, 16, 17

Both the mean and the median estimate where the center of a dataset is located. However, depending on the nature of the data, either the mean or the median may be more useful for describing the center of the dataset.

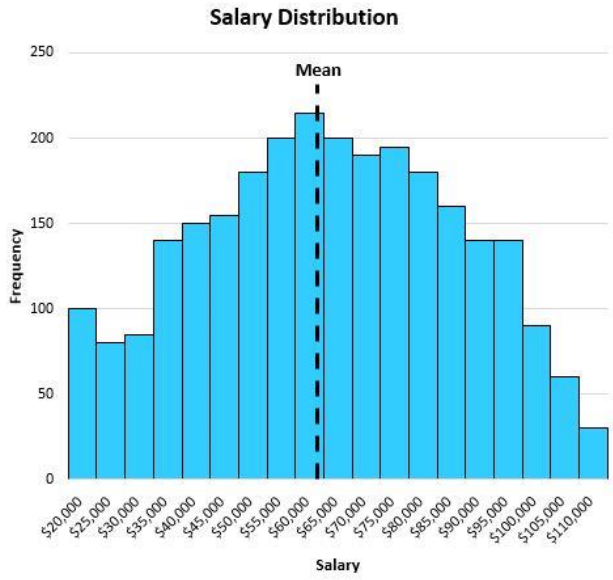
When to Use the Mean

It’s best to use the **mean** to describe the center of a dataset when the distribution is mostly symmetrical and there are no outliers.

For example, suppose we have the following distribution that shows the salaries of residents in a certain city:



Since this distribution is fairly symmetrical (if you split it down the middle, each half would look roughly equal) and there are no outliers, we can use the mean to describe the center of this dataset. The mean turns out to be $63,000, which is located approximately in the center of the distribution:

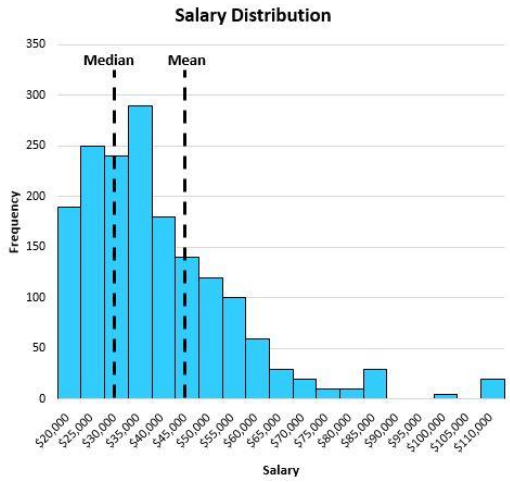


When to Use the Median

It is best to use the median when the distribution is either skewed or there are outliers present.

Skewed Data: For example, consider the following distribution of salaries for residents in a certain city:

When a distribution is skewed, the median does a better job of describing the center of the distribution than the mean

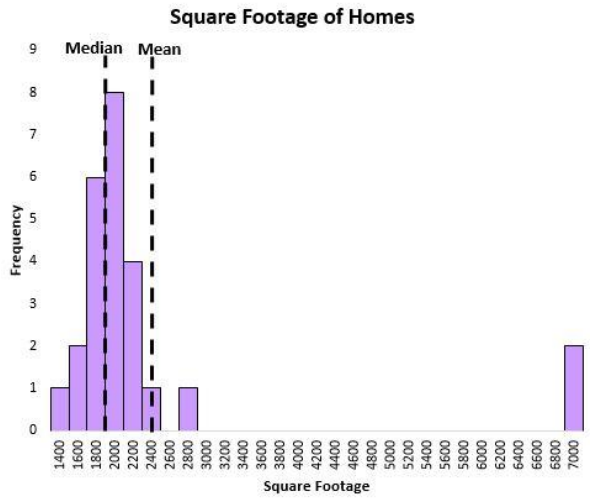


The median does a better job of capturing the "typical" salary of a resident than the mean. This is because the large values on the tail end of the distribution tend to pull the mean away from the center and towards the long tail.

In this example, the mean tells us that the typical individual earns about $47,000 per year while the median tells us that the typical individual only earns about $32,000 per year, which is much more representative of the typical individual.

Outliers:

The median also does a better job of capturing the central location of a distribution when there are outliers present in the data. For example, consider the following chart that shows the square footage of houses on a certain street:



The mean is heavily influenced by a couple extremely large houses, while the median is not. Thus, the median does a better job of capturing the "typical" square footage of a house on this street compared to the mean.

- Both the mean and the median can be used to describe where the "center" of a dataset is located.
- It's best to use the mean when the distribution of the data values is symmetrical and there are no clear outliers.
- It's best to use the median when the the distribution of data values is skewed or when there are clear outliers.

15. The likelihood is **the probability that a particular outcome is observed when the true value of the parameter is , equivalent to the probability mass on** ; it is not a probability density over the parameter . The likelihood should not be confused with , which is the posterior probability of given the data.

The terms Likelihood and Probability are used interchangeably, but few people know the differences between the two.

In layman's terms, the two terms are interchangeable. The terms "likelihood" and "probability" refer to the likelihood of events occurring. In terms of philosophy, the two words have the same denotative meaning. However, these two terms are used in completely different contexts Probability is a branch of mathematics that deals with the possibility of a random experiment occurring. The term "probability" refers to the possibility of something happening.

- The term Likelihood refers to the process of determining the best data distribution given a specific situation in the data.

- When calculating the probability of a given outcome, you assume the model's parameters are reliable.

- However, when you calculate the likelihood, you're attempting to determine whether the parameters in a model can be trusted based on the sample data you have observed.

For example: Suppose you have an unbiased coin. If you flip the coin, the probability of getting head and a tail is equal, which is 0.5

Now suppose the same coin is tossed 50 times, and it shows heads only 14 times. You would assume that the likelihood of the unbiased coin is very low. If the coin were fair, it would have shown heads and tails the same number of times. When calculating the probability of coin getting heads, you assume that P(head) = 0.5

However, when calculating the likelihood, you are trying to find if the model parameter ($p = 0.5$) is correctly specified or not.

The fact that a coin only lands on heads 14 times out of 50 makes you highly suspicious that the true probability of a coin landing on heads on a given toss is $p = 0.5$.

.