STATISTICS WORKSHEET- 6

1.Which of the following can be considered as random variable?

D

2.Which of the following random variable that take on only a countable number of possibilities?

A

3. Which of the following function is associated with a continuous random variable?

A

4.The expected value or _____ of a random variable is the center of its distribution.

C

5. Which of the following of a random variable is not a measure of spread?

A

6. The _____ of the Chi-squared distribution is twice the degrees of freedom

A

7. The beta distribution is the default prior for parameters between ___

C

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?

B

9. Data that summarize all observations in a category are called _____ data.

B

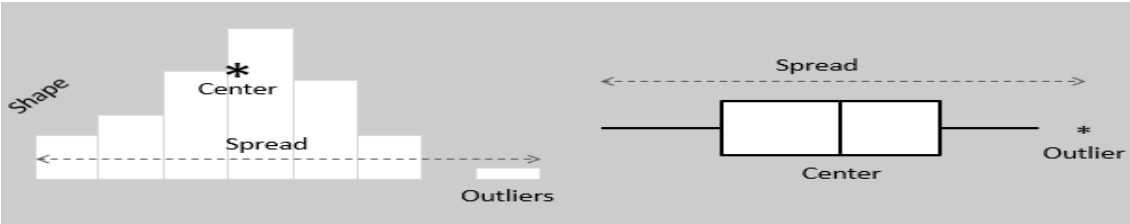
10. What is the difference between a boxplot and histogram?

Ans:

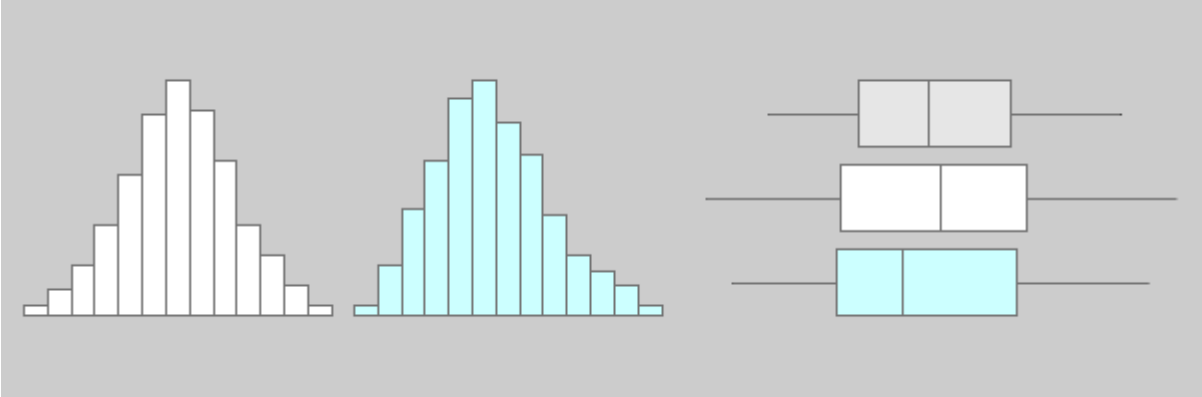
Histograms indicate the whole frequency distribution of a variable, whereas the boxplot summarises its most prominent features. These features include median and spread as well as the extent and nature of departures from symmetry, and the possible presence of observations having extreme values (outliers).

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.

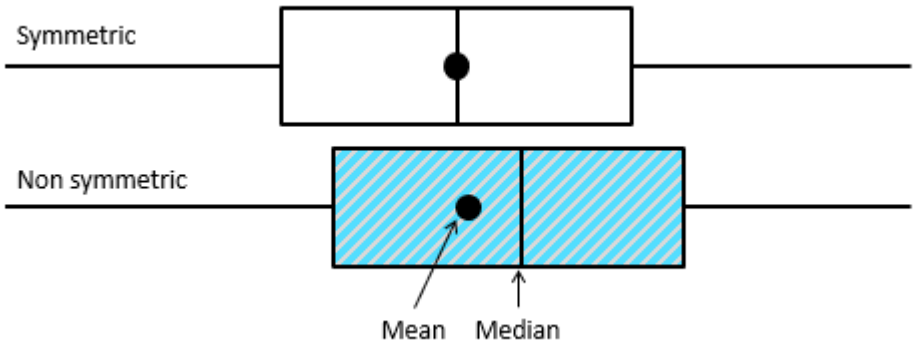
Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.



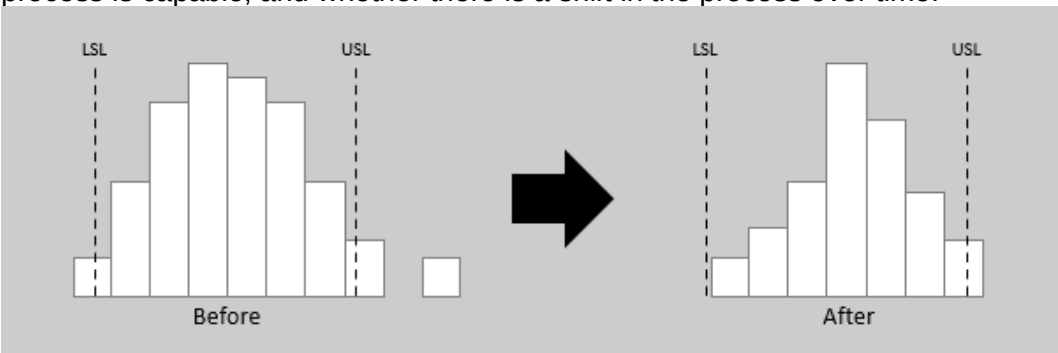
Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed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



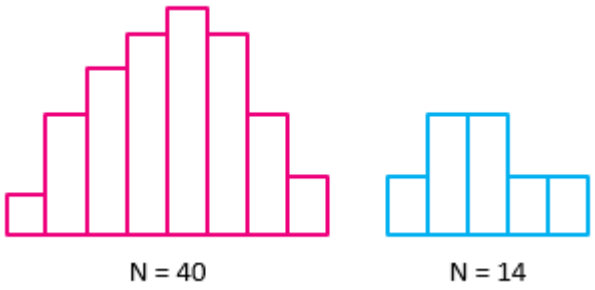
lthough histograms are better in displaying the distribution of data, you can use a box plot to tell if the distribution is symmetric or skewed. In a symmetric distribution, the mean and median are nearly the same, and the two whiskers has almost the same length.

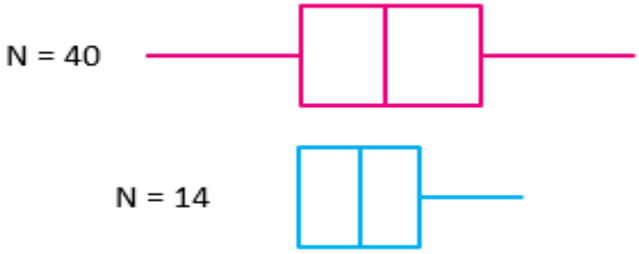


We can use histograms and box plots to verify whether an improvement has been achieved by exploring the data before and after the improvement initiative. Both tools can be helpful to identify whether variability is within specification limits, whether the process is capable, and whether there is a shift in the process over time.

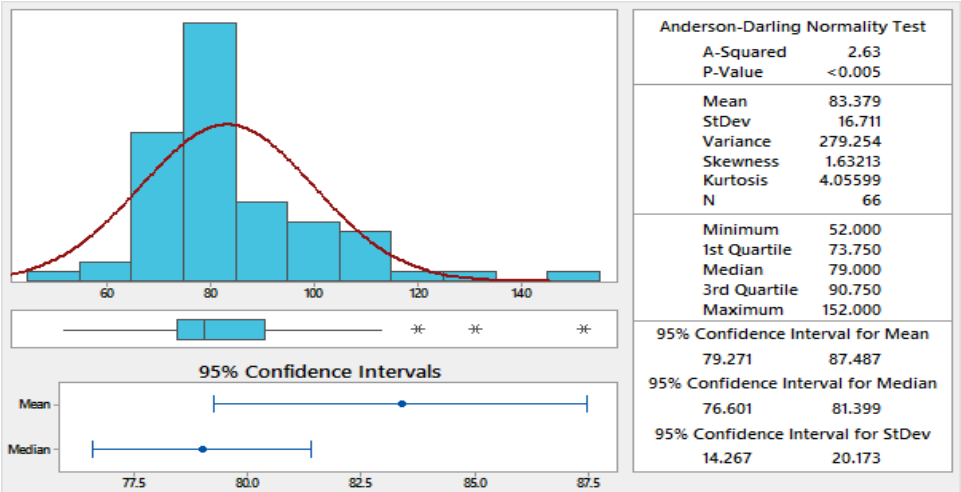


Both histograms and box plots are ideal to represent moderate to large amount of data. They may not accurately display the distribution shape if the data size is too small. In practice, a sample size of at least 30 data values would be sufficient for both tools.





Many statistical applications allow the option of summarizing your data graphically (including plotting the data on histograms and box plots as shown below). This can reveal unusual observations in your data that should be investigated before performing detailed statistical analysis.



Histograms and box plots are very similar in that they both help to visualize and describe numeric data. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space

11. How to select metrics?

Ans:
metric for customer satisfaction. It also had all the characteristics of a badly chosen metric:
No real ownership (read: accountability)
Nearly impossible to root-cause
Home-grown, thus no realistic benchmarking information.
Multiple interpretations (shipped v. delivered, factory v. logistics)
Everybody was impacted by it's poor performance (as it was linked to profit share)
And maybe most important: Unclear value to the customer
A proper metric has an owner, it is meaningful to your customer (whether internal or external) and it has a proper definition. A metric like predictability never really improves and it becomes a curse to anyone that touches it

Use standards. I prefer metrics that have been tested by others;

Measure yourself the way your customer measures you

Only measure metrics that have an owner
The final benefit is the endorsement of approx. 2500 companies across all industries that these are metrics that are valuable to them. If you remember the problems with the predictability metric you can see the benefits of SCOR metrics: No discussion on how it is measured, a clear linkage to processes that may cause the poor performance and endorsements on the value of the metric itself.

	Attribute	Metric (level 1)
Customer	Reliability	Perfect Order Fulfillment
	Responsiveness	Order Fulfillment Cycle Time
	Flexibility	Supply Chain Flexibility
		Supply Chain Adaptability†
Internal	Cost	Supply Chain Management Cost
		Cost of Goods Sold
	Assets	Cash-to-Cash Cycle Time
		Return on Supply Chain Fixed Assets

† upside and downside adaptability metrics

This table shows the SCOR performance attributes (I called them categories) and the highest level metrics. Levels in metrics indicate whether you look at the overall performance or on a detailed aspect of the supply-chain. metrics span the performance of the total supply-chain from procurement of the materials, through producing the product, to delivering and installing the product at your customer. If I want to focus on the customer then I will be directed towards, Reliability, Responsiveness and Flexibility. The importance of one versus the other is determined by industry competitive advantage for your company, or even more importantly how your customer measures you. My company's choice would have been reliability (predictability indicated the accuracy of shipping or sometimes delivering against the day we promised we would ship or deliver.

12. How do you assess the statistical significance of an insight?
Ans:
If a result is **statistically significant**, that means it's unlikely to be explained solely by chance or random factors. In other words, a statistically significant result has a very low chance of occurring if there were no true effect in a research study.
The *p* value, or probability value, tells you the statistical significance of a finding. In most studies, a *p* value of 0.05 or less is considered statistically significant, but this threshold can also be set higher or lower.

A **null hypothesis** (H_0) always predicts no true effect, no relationship between variables, or no difference between groups.
An **alternative hypothesis** (H_a or H_1) states your main prediction of a true effect, a relationship between variables, or a difference between groups.
Hypothesis testing always starts with the assumption that the null hypothesis is true. Using this procedure, you can assess the likelihood (probability) of obtaining your results under this assumption. Based on the outcome of the test, you can reject or retain the null hypothesis.

statistical significance: Researchers classify results as statistically significant or non-significant using a conventional threshold that lacks any theoretical or practical basis. This means that even a tiny 0.001 decrease in a p value can convert a research finding from statistically non-significant to significant with almost no real change in the effect. On its own, statistical significance may also be misleading because it's affected by sample size. In extremely large samples, you're more likely to obtain statistically significant results, even if the effect is actually small or negligible in the real world. This means that small effects are often exaggerated if they meet the significance threshold, while interesting results are ignored when they fall short of meeting the threshold.

The strong emphasis on statistical significance has led to a serious publication bias and replication crisis in the social sciences and medicine over the last few decades. Results are usually only published in academic journals if they show statistically significant results—but statistically significant results often can't be reproduced in high quality replication studies.

You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal
Ans:
Many random variables have distributions that are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the

distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.
For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.
What generally never becomes Gaussian, however, is the Uniform Distribution. A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.
Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.
The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.

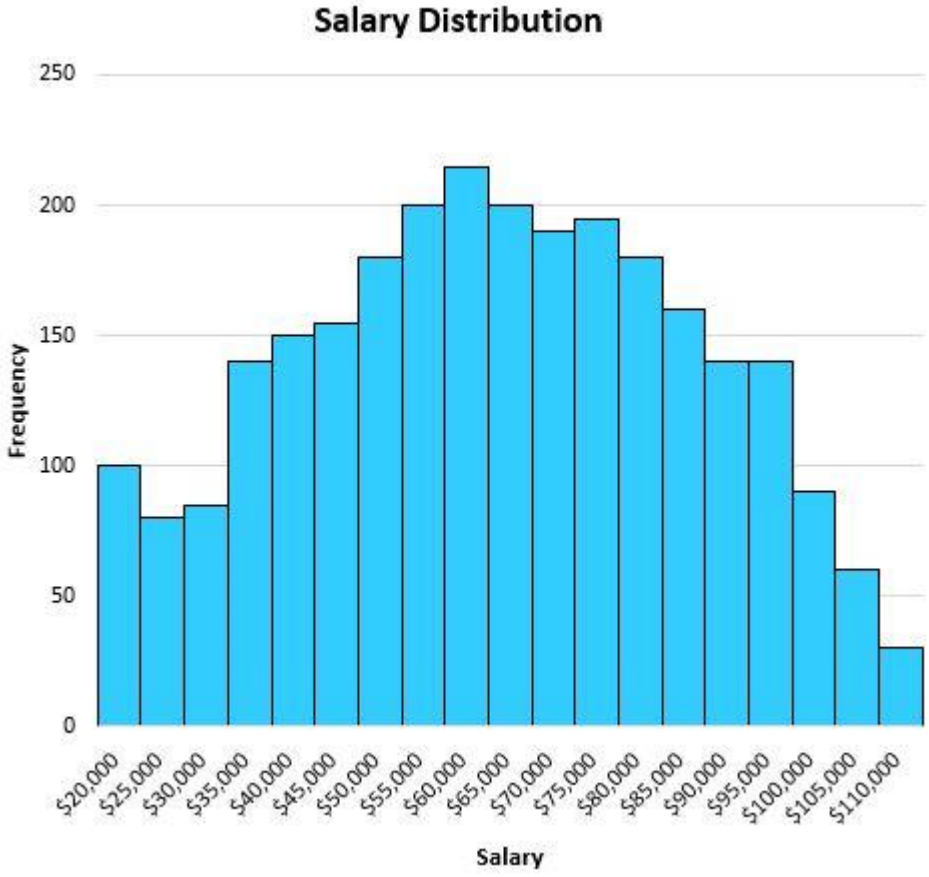
14. Give an example where the median is a better measure than the mean.
Ans:
When there are a number of outliers that positively or negatively skew the data

he 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.

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

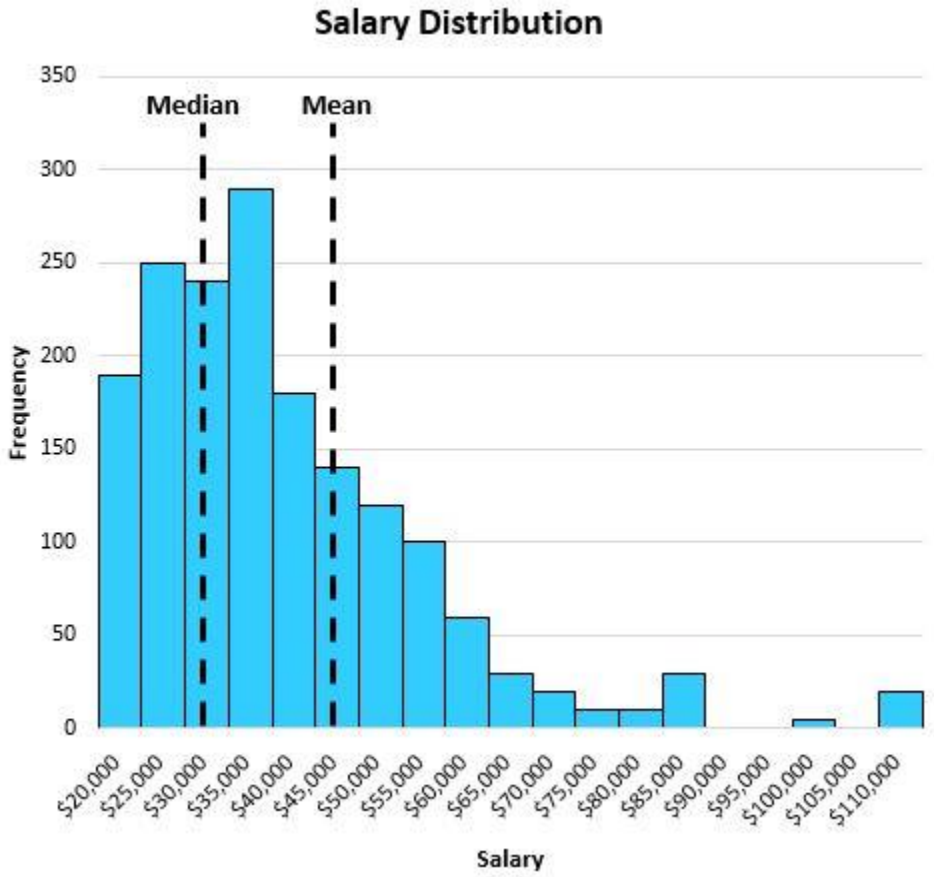


ince this distribution is fairly symmetrical (if you split it down the middle, each half would look roughly equal) and there are no outliers, we can use the mean to describe the center of this dataset.
The mean turns out to be $63,000, which is located approximately in the center of the distribution:

Use the Median:

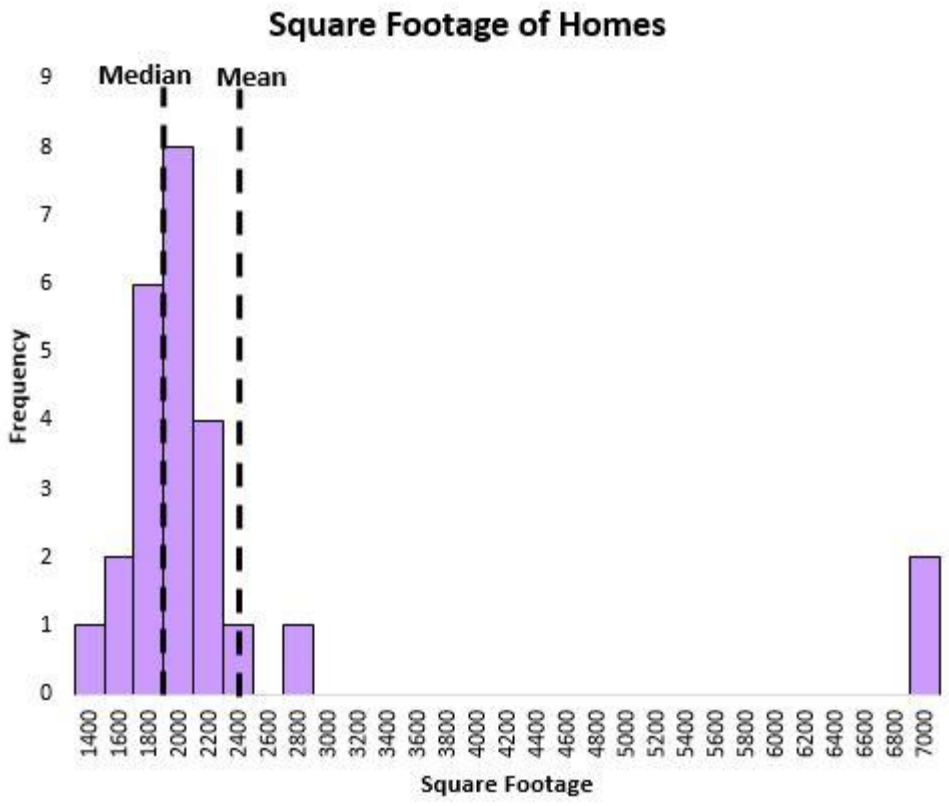
It is best to use the median when the distribution is either skewed or there are outliers present.

Skewed Data:
When a distribution is skewed, the median does a better job of describing the center of the distribution than the mean.
For example, consider the following distribution of salaries for residents in a certain city



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.

15. What is the Likelihood?

Ans:

In the field of statistics, researchers are interested in making inferences from data. The data is collected from a population; the data drawn from a population is called a sample. In a statistical experiment we consider taking a data sample from some infinite population, where each sample member/ unit is associated with an observed value of some variable. In frequentist statistics, the concept of an infinite population is adopted so we can assume that observations may be drawn repeatedly without limit.

Once we have a particular data sample, experiments can be performed to make inferences about features about the population from which a given data sample is drawn. Fundamentally speaking, the feature of a population that a researcher is interested in making inferences about is called a parameter. In frequentist statistics a parameter is never observed and is estimated by a probability model.

A fundamental role in the theory of statistical inference is played by the likelihood function. Given a particular vector of observed values x, the likelihood function

$L(\theta;x)$ is the joint probability density function $f(x;\theta)$ but the change in notation considers the pdf as a function of the parameter θ. Hence
$L(\theta;x)=f(x;\theta)$

he likelihood function is an expression of the relative likelihood of the various possible values of the parameter θ which could have given rise to the observed vector of observations xx. Given a statistical model, we are comparing how good an explanation the different values of θ provide for the observed data we see x. In other words, given that we observe some data, what is the probability distribution which is most likely to have given rise to the data that we observe? Often it will be useful to speak about the likelihood function $L(\theta;x)$ and its logarithm – the log likelihood function $l=\ln(L(\theta;x))$