STATISTICS

1. Which of the following can be considered as random variable?
   Answer – d
2. Which of the following random variable that take on only a countable number of possibilities?
   Answer – a
3. Which of the following function is associated with a continuous random variable?
   Answer – a
4. The expected value or _____ of a random variable is the center of its distribution.
   Answer – c
5. Which of the following of a random variable is not a measure of spread?
   Answer –c
6. The _____ of the Chi-squared distribution is twice the degrees of freedom.
   Answer – a
7. The beta distribution is the default prior for parameters between _____
   Answer – c
8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?
   Answer – b
9. Data that summarize all observations in a category are called _____ data.
   Answer – b
10. What is the difference between a boxplot and histogram?
    Answer –Histograms are great for showing what data ranges are most and least common, but they do not tell details like the range or the median. You can use box plots to present these values. They have 5 vertical lines. The lines farthest on the left and right tell the least and greatest values of the data set. The line in the middle is the median. The other two lines are called the lower quartile and upper quartile. The lower quartile line is on the left of the median, and it tells us that one-quarter of the data points are less than or equal to the lower quartile. The upper quartile is on the right of the median and tells us that one-quarter of the data points are greatest than or equal to the upper quartile.

    The wait times for a rollercoaster, in minutes, are: 51, 54, 55, 56, 57, 57, 58, 58, 58, 59, 59, 59, 59, 59, 60, 61, 61, 61, 61, 62, 62, 64, 64, 66, 67, 69, 70, 71, 71. To start making the box plot, locate the least and greatest values. Luckily, this data set is already sorted from least to greatest, so you can see that these values are 51 and 71. Place and label these values on the plot. Next, find the median. The number in the middle is 60. That means half of the people waited greater than or equal to 60 minutes, and half the people waited less than or equal to 60 minutes. To find the lower quartile, we find the time that is the middle data point in the range 51 to 60, and to find the upper quartile we find the middle data point in the range 60 to 71. The lower quartile is 58 and the upper quartile is 64. If 25% of people waited 58 minutes or less, and 25% of people waited 64 minutes or more, that means that at least half of the people waited between 58 and 64 minutes. This is called the interquartile range.

    If a data set has a lot of different measurements, displaying it using line graphs does not always help to interpret the information. Instead, you can display the data using a type of bar graph called a histogram. In a histogram, bars represent ranges instead of individual values. These bars are called bins, and they are presented continuously with no spaces

between them. Each bar represents a range of data points, and the height of the bar tells us how many data points are in that range

When you have a lot of data, you first have to decide how many bins you would like to use, and what the range of each bin should be. In a bike race, the distances in kilometers that cyclists rode are: 5, 8.25, 15.5, 18, 20, 22.5, 28, 28, 29.5, 30, 30, 36.5, 38, 42.5, 45.75, 46, 47, 48, 48, 50, 50, 52, 55, 58, 58, 59, 63.25, 65.5, 67, 70, 70, 72, 75, 75, 76, 83, 87.75, 94.5, 95. You can organize the data into 5 bins, the first one 0-19, then 20-39, 40-59, 60-79, and 80-99. You can then draw the axes for the graph and label the bin sizes at the bottom along the x-axis, and label it "Distance Biked." The y-axis can be labeled "Number of Cyclists." Now as you read the data, you can make a tick to count each time a point contributes to a bin. The 0-19 bin has a frequency of 4, 20-39 has 9, 40-59 has 13, 60-79 has 9, and 80-99 has 4. You can see that the most common distance biked is 40-59 kilometers! If you choose bin size 5, you would have a lot more bars and they would be harder to interpret. If you choose bin size 40, you would only have 3 bars, which is not clear either. It is important to choose a bin size that helps you make sense of the data.

11. How to select metrics?

Answer – In a previous life I was in charge of management reporting for the pan-European logistics organization for Compaq. My team was responsible for monitoring and reporting the daily, weekly, monthly and quarterly performance and progress of the operations. We provided reports on shipments-to-date, invoices send, estimated shipments for the remainder of the day, week or month, order cycle time and average shipment lead-times, and on and on. None of these metrics would keep me awake at night or would cause heavy discussion on accuracy and validity of the metric. That was reserved for one metric: "Predictability".

Predictability was our #1 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
  Recognize this type of metric?

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. (My nick-name was 'mister predictability').

Nowadays I take a different approach on metrics. I use three basic rules in selecting metrics:

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

2. Measure yourself the way your customer measures you

3. Only measure metrics that have an owner

So, what is this 'standards thing' you may ask. Let's look at that logistics operation in Europe: clearly a supply-chain function with warehousing, consolidation and shipping as core activities. The recommend standard supply-chain reference framework with metrics is SCOR. I use SCOR whenever I need to look at the performance of a supply-chain. It provides me with 3 valuable things in the metric space:

- Standard metrics. These are pre-defined metrics; how is it measured
- Metric categorization. Metrics are grouped together based on the type of performance they represent (e.g. Reliability, flexibility, cost, etc).
- And very important a linkage to those processes that influence a metric's performance.

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

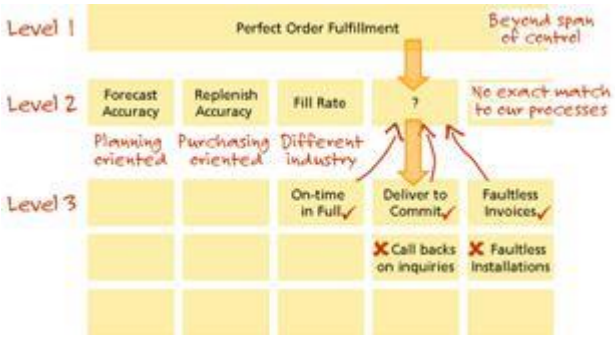
This table shows the SCOR performance attributes (I called them categories) and the highest level metrics. Levels in metrics indicate whether you look at the overall performance or on a detailed aspect of the supply-chain.

SCOR = Supply Chain Operations Reference, maintained by the Supply-Chain Council,These level-1 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).

What to do if company is not so standard? I recently worked with a military organization responsible for maintaining a fleet. Their core processes are maintenance, repair and overhaul to keep their fleet operational. In the SCOR language this is covered in the Returns processes. They had been instructed to "use the same metrics as the great supply-chain examples -like WalMart- do". But however they looked at their operation the inventory metrics did not seem to make sense. My question for them was "How do your customers truly measure you?" The answer: The cost per flying hour. We found this to be very similar

to the 'return on supply chain assets' metric in SCOR. My recommendation: Adapt the SCOR metrics to your organization. Replace a standard SCOR metric with your similar metric. But make sure you retain the linkage to the processes. This way you can root-cause the issues and design a permanent solution to your performance gaps later. Oh, and forget about 'the WalMart metrics' and start looking at how companies that operate similar processes to your business measure themselves. As an example: transportation companies maintain fleets (see, air and road).

How do you find the metric at the right level for your organization? If the metric is too high level it measures beyond the boundaries of your organization. The way to find the right level is to drill down until you find the right metric for your organization.



In this example, no level 2 metric is available that matches the need. Yet level-3 was too detailed. Therefore I consolidated the appropriate level-3 metrics to a new level 2 that does match my organization. By consolidation of level-3 metrics I retain the links to the processes (via level 3). Furthermore I already have some metrics that I can link to the departments within my logistics organization (faultless invoices for my billing department).

And that's how I select a metric. 3 Simple checks make all the difference:

[ ] This is a standard metric (If not replace with a standard metric or start with standard lists,

[ ] This is how my customer measures me (If not don't use it or if not exactly find the closest match)

[ ] This metric has an owner (If not decompose the metric the next level and test again, continue until all components have owners)

12. How do you assess the statistical significance of an insight?
    Answer –To assess statistical significance, you would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, you'd 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.

13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

Answer –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.

Answer — Median :-

The median is the middle value. It is the value that splits the dataset in half, making it a natural measure of central tendency.

To find the median, order your data from smallest to largest, and then find the data point that has an equal number of values above it and below it. The method for locating the median varies slightly depending on whether your dataset has an even or odd number of values. I'll show you how to find the median for both cases. In the examples below, I use whole numbers for simplicity, but you can have decimal places.

In the dataset with the odd number of observations, notice how the number 12 has six values above it and six below it. Therefore, 12 is the median of this dataset.

Median Odd
23
21
18
16
15
13
12
10
9
7
6
5
2

When there is an even number of values, you count in to the two innermost values and then take the average. The average of 27 and 29 is 28. Consequently, 28 is the median of this dataset.

Median Even
40
38
35
33
32
30
29
28
27
26
24
23
22
19
17

Outliers and skewed data have a smaller effect on the mean vs median as measures of central tendency. To understand why, imagine we have the Median dataset below and find that the median is 46. However, we discover data entry errors and need to change four values, which are shaded in the Median Fixed dataset. We'll make them all significantly higher so that we now have a skewed distribution with large outliers.

Median	Median Fixed
69	112
56	93
54	89
52	82
47	47
46	46
46	46
45	45
43	43
36	36
35	35
34	34
31	31

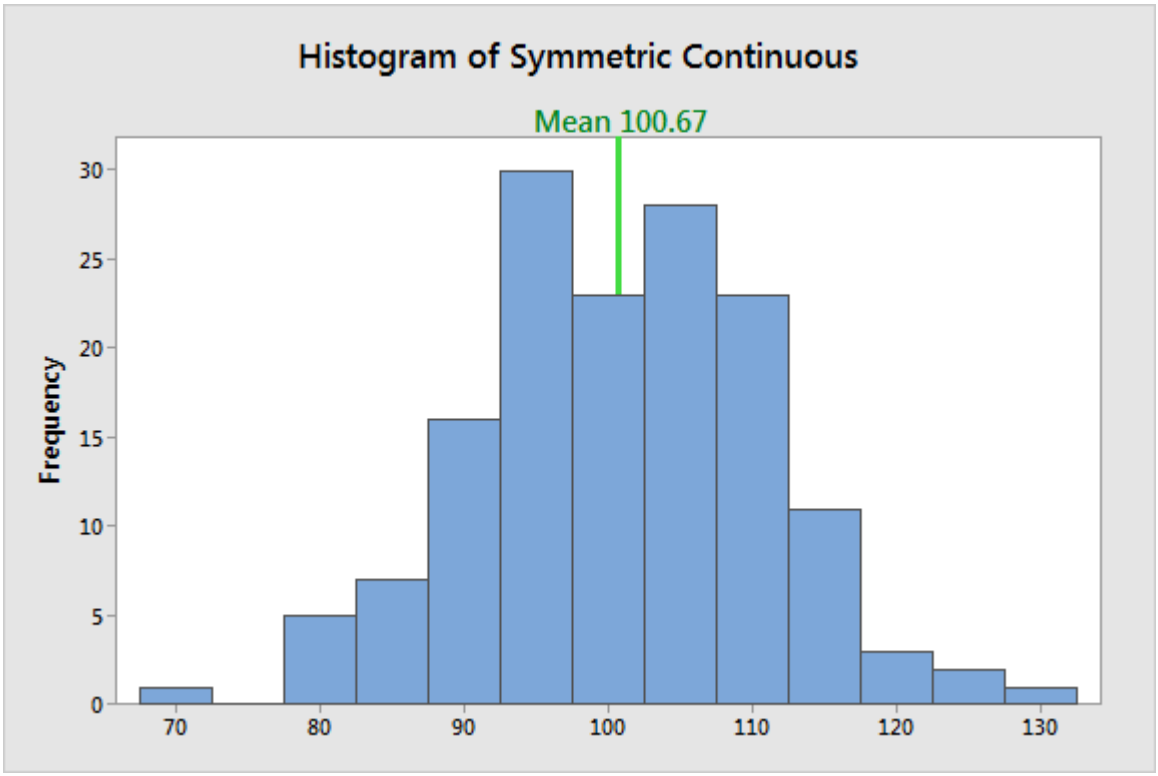
As you can see, the median doesn't change at all. It is still 46. When comparing the mean vs median, the mean depends on all values in the dataset while the median does not. Consequently, when some of the values are more extreme, the effect on the median is smaller. Of course, with other types of changes, the median can change. When you have a skewed distribution, the median is a better measure of central tendency than the mean.

Mean :-

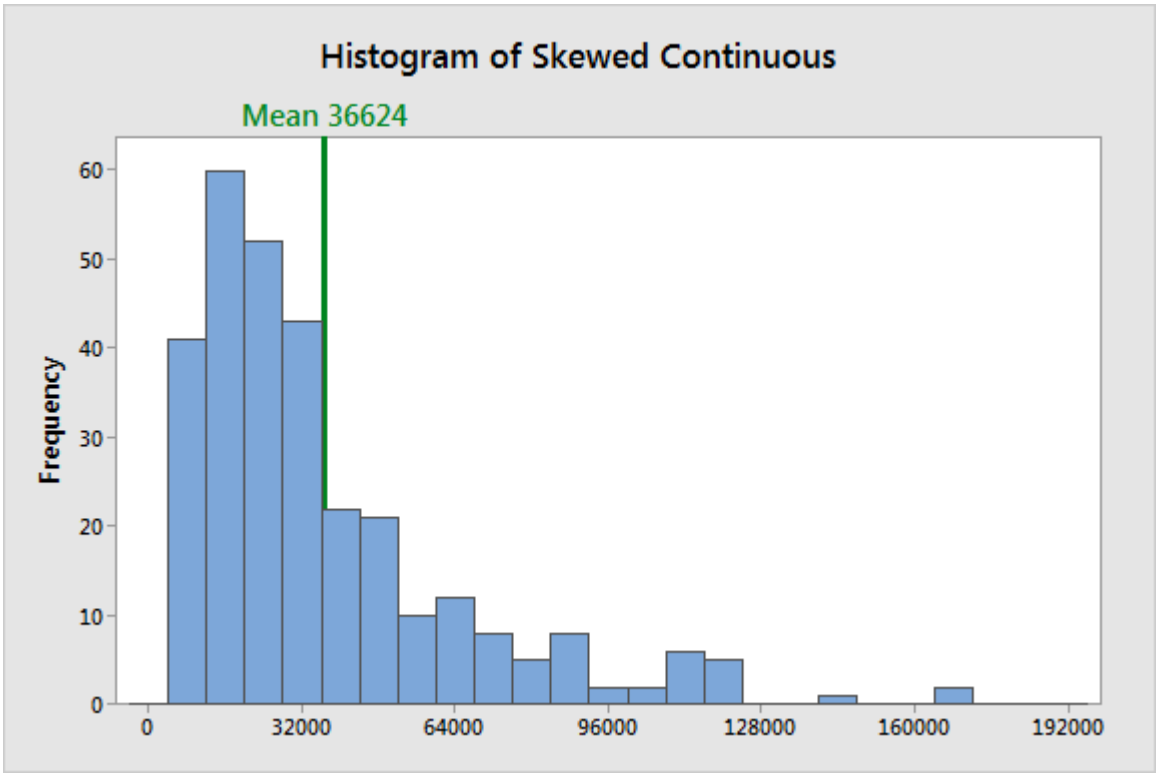
The mean is the arithmetic average, and it is probably the measure of central tendency that you are most familiar. Calculating the mean is very simple. You just add up all of the values and divide by the number of observations in your dataset.

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

The calculation of the mean incorporates all values in the data. If you change any value, the mean changes. However, the mean doesn't always locate the center of the data accurately. Observe the histograms below where I display the mean in the distributions.



In a symmetric distribution, the mean locates the center accurately.



However, in a skewed distribution, the mean can miss the mark. In the histogram above, it is starting to fall outside the central area. This problem occurs because outliers have a substantial impact on the mean as a measure of central tendency. Extreme values in an

extended tail pull the mean away from the center. As the distribution becomes more skewed, the mean is drawn further away from the center. Consequently, it's best to use the mean as a measure of the central tendency when you have a symmetric distribution. More about this issue when we look at the mean vs median!

In statistics, we generally use the arithmetic mean, which is the type I discuss in this post. However, there are other types of means, such as the geometric mean. Read my post about the geometric mean to learn when it is a better measure. Use a weighted mean when you need to place differing importance on the values.

15. What is the Likelihood?

Answer – In parametric models like linear regression and logistic regression, we are given a set of data points with the goal of finding the parameters of these models that best fit the observed data. Let's consider the same house price example that we introduced in the previous section. We want to fit some statistical model to predict house prices given some information about the property such as number of bedrooms in the house, size of house in square footage (sqft), and age of the house.

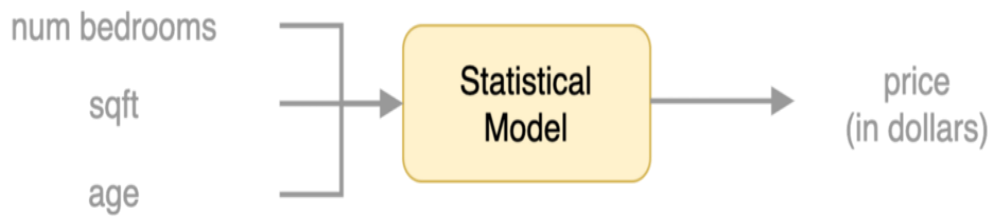


Figure 1

These are 3 features that go into the model here, but there can be more in practice. Let us assume we want to perform a linear regression. Because of this assumption, the input features and output label are related in the following way.

$$\begin{aligned} price = \theta_3 \times \ & num\ bedrooms \\ + \theta_2 \times \ & sqft \\ + \theta_1 \times \ & age \\ + \theta_0 + \epsilon & \end{aligned}$$

Equation 1

One can write this more generally with the following form.

$$y_i = \theta_3 x_{i3} + \theta_2 x_{i2} + \theta_1 x_{i1} + \theta_0 + \epsilon$$

Equation 2

The *x* terms are the features of the i^{th} house, *y* is the price of the i^{th} house, the θ terms are the coefficients of each feature, and the epsilon ϵ denotes an irreducible error. It is error from inherent system randomness and also occurs because some features are not accounted for.

To construct this linear regression model, we need to know the values of the θ terms. To find the θ terms, we need examples of house features and their prices. i.e we need pairs of (x, y) to fill the values in *Equation 3* to estimate the θ terms. This is why we need training data.

Remember our imaginary city Databerg ? Let's add details to make this data useful for training a model. We have access to *10,000* house records in Databerg. Each record has information about a house: number of bedrooms; size of the house in square feet; age of the house; and the price at which this house is evaluated. Since we want to predict the price of a house, the label *y* is this price. The other fields in this imaginary dataset are the features *x* that are inputs to our linear regression to predict the corresponding price. This training data looks like the table below. One house is *$757,000*, a second house is *$780,000* a third house is *$680,000,* and so on.

	x			y
index	num bedrooms	sqft	age	price
1	3	3,024	10	$757,000
2	4	3,225	5	$780,000
3	3	2,200	7	$680,000
- - -	- - -	- - -	- - -	- - -
10,000	1	1,200	4	$890,000

Figure 2

If we assume the values of the θ terms, we can quantify how well the linear regression model fits training data using the likelihood function. In the end, we want to determine the θ terms that will *best fit* the given data; in other words, we want to determine the value of the θ terms that will maximize the likelihood function. This is translated into math as follows.

$$\theta_0^{MLE}, \theta_1^{MLE}, \theta_2^{MLE}, \theta_3^{MLE} = \arg \max_{\theta_0, \theta_1, \theta_2, \theta_3} \mathcal{L}(\theta_0, \theta_1, \theta_2, \theta_3)$$

Equation 3

We will explain these terms shortly. But before that, let us get rid of this cumbersome notation by representing all the θ terms in a vector form.

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

Equation 4

Now, *Equation 4* can be written in the more general and concise form.

$$\hat{\theta}^{MLE} = \arg \max_{\theta} \mathcal{L}(\theta)$$

Equation 5

This notation is telling us a few things: the likelihood function *L* is a function of the model parameters θ; the *arg max* function returns the value of θ that maximizes this likelihood function *L*. Quite literally by definition, this value of θ obtained is the *maximum likelihood estimation* of θ. To distinguish the variable θ used on the right hand side from this specific value of θ we seek on the left, we add a *MLE* superscript to the latter. Furthermore, the hat on the θ maximum likelihood estimate shows this is value is just that — an estimate.