WORKSHEET

STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

answer: a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

answer: a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

answer: b) Modeling bounded count data

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

answer: c) The square of a standard normal random variable follows what is called chi-squared distribution

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

answer: c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

answer: b) False

7. 1. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

answer: b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

b) 0

9. Which of the following statement is incorrect with respect to outliers?

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

answer: c) Outliers cannot conform to the regression relationship

WORKSHEET

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

A normal distribution is **an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme**.

- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
- In graphical form, the normal distribution appears as a "bell curve".
- The normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- Many naturally-occurring phenomena tend to approximate the normal distribution.
- In finance, most pricing distributions are not, however, perfectly normal.

normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation.

The normal distribution model is important in statistics and is key to the Central Limit Theorem (CLT). This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance).
The normal distribution is one type of symmetrical distribution. Symmetrical distributions occur when where a dividing line produces two mirror images. Not all symmetrical distributions are normal, since some data could appear as two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

Properties of the Normal Distribution
The normal distribution has several key features and properties that define it.

First, its mean (average), median (midpoint), and mode (most frequent observation) are all equal to one another. Moreover, these values all represent the peak, or highest point, of the distribution. The distribution then falls symmetrically around the mean, the width of which is defined by the standard deviation.

All normal distributions can be described by just two parameters: the mean and the standard deviation.

The Empirical Rule
For all normal distributions, 68.2% of the observations will appear within plus or minus one standard deviation of the mean; 95.4% of the observations will fall within +/- two standard deviations; and 99.7% within +/- three

standard deviations. This fact is sometimes referred to as the "empirical rule," a heuristic that describes where most of the data in a normal distribution will appear.

This means that data falling outside of three standard deviations ("3-sigma") would signify rare occurrences.

Skewness
Skewness measures the degree of symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero.

If the distribution of a data set instead has a skewness less than zero, or negative skewness (left-skewness), then the left tail of the distribution is longer than the right tail; positive skewness (right-skewness) implies that the right tail of the distribution is longer than the left.

Kurtosis
Kurtosis measures the thickness of the tail ends of a distribution in relation to the tails of a distribution. The normal distribution has a kurtosis equal to 3.0.

Distributions with larger kurtosis greater than 3.0 exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean). This excess kurtosis is known in statistics as leptokurtic, but is more colloquially known as "fat tails." The occurrence of fat tails in financial markets describes what is known as tail risk.

Distributions with low kurtosis less than 3.0 (platykurtic) exhibit tails that are generally less extreme ("skinnier") than the tails of the normal distribution.

The Formula for the Normal Distribution
The normal distribution follows the following formula. Note that only the values of the mean (μ) and standard deviation (σ) are necessary
Normal Distribution Formula.

where:

x = value of the variable or data being examined and f(x) the probability function
μ = the mean
σ = the standard deviation
How Normal Distribution Is Used in Finance
The assumption of a normal distribution is applied to asset prices as well as price action. Traders may plot price points over time to fit recent price action into a normal distribution. The further price action moves from the mean, in this case, the greater the likelihood that an asset is being over or undervalued. Traders can use the standard deviations to suggest potential trades. This type of trading is generally done on very short time frames as larger timescales make it much harder to pick entry and exit points.

Similarly, many statistical theories attempt to model asset prices under the assumption that they follow a normal distribution. In reality, price distributions tend to have fat tails and, therefore, have kurtosis greater than three. Such assets have had price movements greater than three standard deviations beyond the mean more often than would be expected under the assumption of a normal distribution. Even if an asset has gone through a long period where it fits a normal distribution, there is no guarantee that the past performance truly informs the future prospects.

Example of a Normal Distribution

Many naturally-occurring phenomena appear to be normally-distributed. Take, for example, the distribution of the heights of human beings. The average height is found to be roughly 175 cm (5' 9"), counting both males and females.

As the chart below shows, most people conform to that average. Meanwhile, taller and shorter people exist, but with decreasing frequency in the population. According to the empirical rule, 99.7% of all people will fall with +/- three standard deviations of the mean, or between 154 cm (5' 0") and 196 cm (6' 5"). Those taller and shorter than this would be quite rare (just 0.15% of the population each).

Heights
What Is Meant By the Normal Distribution?
The normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation. It is visually depicted as the "bell curve."

Why Is the Normal Distribution Called "Normal?"
The normal distribution is technically known as the Gaussian distribution, however it took on the terminology "normal" following scientific publications in the 19th century showing that many natural phenomena appeared to "deviate normally" from the mean. This idea of "normal variability" was made popular as the "normal curve" by the naturalist Sir Francis Galton in his 1889 work, Natural Inheritance.

What Are the Limitations of the Normal Distribution in Finance?
Although the normal distribution is an extremely important statistical concept, its applications in finance can be limited because financial phenomena—such as expected stock-market returns—do not fall neatly within a normal distribution. In fact, prices tend to follow more of a log-normal distribution that is right-skewed and with fatter tails. Therefore, relying too heavily on a bell curve when making predictions about these events can lead to unreliable results. Although most analysts are well aware of this limitation, it is relatively difficult to overcome this shortcoming because it is often unclear which statistical distribution to use as an alternative.
normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation.

The normal distribution model is important in statistics and is key to the Example of multiple linear regression

Linear regression with a single predictor variable is known as *simple regression.* In real-world applications, there is typically more than one predictor variable. Such regressions are called *multiple regression.* For more information, check out this post on why you should not use multiple linear regression for Key Driver Analysis with example data for multiple linear regression examples.
Returning to the Benetton example, we can include **year** variable in the regression, which gives the result that **Sales = 323 + 14 Advertising + 47 Year.** The interpretation of this equation is that every extra million Euro of advertising expenditure will lead to an extra 14 million Euro of sales and that sales will grow due to non-advertising factors by 47 million Euro per year.

Checking the quality of regression models

Estimating a regression is a relatively simple thing. The hard bit of using regression is avoiding using a regression that is wrong. Below are standard regression diagnostics for the earlier regression.

The column labelled **Estimate** shows the values used in the equations before. These estimates are also known as the *coefficients* and *parameters.* The **Standard Error** column quantifies the uncertainty of the estimates. The

standard error for Advertising is relatively small compared to the Estimate, which tells us that the Estimate is quite precise, as is also indicated by the high ***t*** *(*which is **Estimate / Standard**), and the small ***p***-value. Furthermore, the R-Squared statistic of 0.98 is very high, suggesting it is a good model.

Linear Regression: Sales

	Estimate	Standard Error	*t*	*p*
(Intercept)	167.68	58.94	**2.85**	**.025**
Advertising	23.42	1.37	**17.13**	**< .001**

n = 9 cases used in estimation; R-squared: 0.9767; Correct predictions: 88.89%; AIC: 100.34; multiple comparisons correction: None

A key assumption of linear regression is that all the relevant variables are included in the analysis. We can see the importance of this assumption by looking at what happens when **Year** is included. Not only has Advertising become much less important (with its coefficient reduced from 23 to 14), but the standard error has ballooned. The coefficient is no longer statistically significant (i.e., the *p*-value of 0.22 is above the standard cutoff of .05). This means is that although the estimate of the effect of advertising is 14, we cannot be confident that the true effect is not zero.

Linear Regression: Sales

	Estimate	Standard Error	*t*	*p*
(Intercept)	323.54	177.60	1.82	.118
Advertising	13.99	10.22	1.37	.220
Year	46.60	50.03	0.93	.388

n = 9 cases used in estimation; R-squared: 0.9797; Correct predictions: 88.89%; AIC: 101.13; multiple comparisons correction: None

In addition to reviewing the statistics shown in the table above, there are a series of more technical diagnostics that need to be reviewed when checking regression models, including checking for *outlier*

15. What are the various branches of statistics?

There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

Statistics is all about the interpretation of data.

Descriptive Statistics. In this type of statistics, the data is summarised through the given observations. ...

Inferential Statistics. This type of statistics is used to interpret the meaning of Descriptive statistics. ...

Statistics Example.

29%, Liberal Democrat 23%, Others 12%). This is an example of descriptive statistics – 'describing' or 'summarising' the overall data for people to understand.

Inferential statistics

Inferential statistics is the aspect that deals with making conclusions about the data. This is quite a wide area; essentially you are asking 'What is this data telling us, and what should we do?'

For example, a council might be considering altering the speed limit on a main road, after a number of accidents. They might do this by for example, a number of cars are driving too fast). Note, though, that this may not be the case; everyone might be driving at a perfectly acceptable speed, and the accidents are down to something other than speed (a blind spot or a pothole, for example). This is inferential statistics: take the data you have and make an 'inference' or 'conclusion' from it. We shall see much more of this later when we discuss things such as hypothesis testing, where we test to see whether the data supports a belief that we have.

Discrete and continuous data

Data comes in two distinct types. Discrete data can take distinct values, which can be clearly identified and separated. An example of this is the score obtained by rolling a die, which can only take values of 1, 2, 3, 4, 5 or 6, with nothing in between, and all the scores can be distinguished. By contrast, continuous data can take any value. For example, when you measure the speed of a car, it could take any value, depending on how accurately you measure it – for example 31.2 or 48.28, or 48.281 – basically any value.

A good illustration is to consider the whole numbers (1, 2, 3, etc.), which are clearly distinct from each other (and so discrete), and

the positive real numbers (every number you can think of, including decimals), which are continuous. If you draw a line of the whole numbers between 1 and 10 it looks something like

1 – 2 – 3 – 4 – 5 – 6 – 7 – 8 – 9 – 10

and it is clear that the numbers are distinct. But if you draw a line of the real numbers between 1 and 10, it's just a single straight line – instead of taking one of the 10 discrete values as above, you can now take any value in the range, for example 1.75, 2.895, 9.238984 and so on.

Continuous data will almost always be approximated (to 1 decimal place say, as in the speed of a car, for example), whereas discrete data will be exact (the score obtained by a single dart, for example).

Frequency distributions

Sometimes the actual collection of data isn't very meaningful, and we wish to put the data into 'categories'. Take for example a list of student marks as percentages:

32, 78, 37, 65, 90, 87, 12, 41, 0, 91, 17, 65, 41, 45, 69, 54, 82, 65, 60, 51, 21, 37, 28, 53, 42, 48, 9, 71

Just looking at this doesn't really tell us much. It would be more useful to categorise the students into degree classifications. The fairly standard classifications in universities are the following:

First class: 70 or more

Upper second class: between 60 and 69

Lower second class: between 50 and 59

Third class: between 40 and 49

Fail: less than 40

If you work through the list of students and count how many students fall into each class, you should get the following:

First class: 6

Upper second class: 5

Lower second class: 3

Third class: 5

Fail: 9

Make sure you do this yourself. In lecture notes and books there are often many examples. If you don't get the answer you expected, then either you made a mistake (so do it again), or you didn't understand (so ask), or the lecturer/author made a mistake (this is possible, we all make mistakes): ask, and if they did, they will correct it – and they would like to know!

This is an example of a frequency distribution; instead of allocating the precise mark to each student, you are placing them in an appropriate category to get a more concise view of the results. consider its value to be the midpoint of the range when we come to analyse it: this is something we shall return to later.

The categories are easy to create when you have discrete variables like the student marks, but it gets trickier with continuous variables, because you can have values that lie right on the edge of an interval. For example, you might want to class the speed of cars on a road into the following categories:

less than 30 mph

between 30 mph and 40 mph

between 40 mph and 50 mph

above 50 mph

The problem comes when a car is clocked exactly at one of the boundaries. Where do we put a car that is measured at exactly 40 mph? Should it go in the second or the third category? One obvious solution is simply to make the second category between 30 and 39 mph and the third category between 40 and 49 mph, etc. But then suppose we are measuring our speeds to 1 decimal place. Where does a car clocked at 39.5 mph fit? It goes into no categories at all now!

The solution here is to be more specific in what you mean by the

intervals. Instead of ‘between 30 mph and 40 mph’ you could specify ‘in the range from 30 mph to less than 40 mph’, which you might abbreviate as ‘30 to 40’ using the ‘less than’ symbol . Writing our categories like this, we can define them as

30

30 to 40

40 to 50

50

(note that we need the symbol for ‘greater than or equal to’ in the last one, to make sure that 50 is included in the range) and now every observation will fit into exactly one category.

You could write this in a different way; it’s not really important how it is written as long as it is clear, and you have ensured that every value fits into exactly one category.

; virtually any subject will need some element of data analysis and study. Remember that there are various aspects to statistics: the actual data collection, the presentation of the data (descriptive statistics), and the conclusions that can be drawn (inferential statistics).