How to calculate the prediction interval and why does the formula work?

Imagine the following scenario: You used some method (not relevant here) to forecast the future values of the dataset that is of interest to you. The method, or the model you used, produced the values that back-fit you actual data and you were able to calculate the confidence interval that surrounds your historical model data. You extrapolate your model results into the future and need to define the prediction interval. You should expect that the prediction interval should get wider the further into the future you go, despite the fact that you will maintain the same level of confidence. Graphically, it should look something like Figure 1:

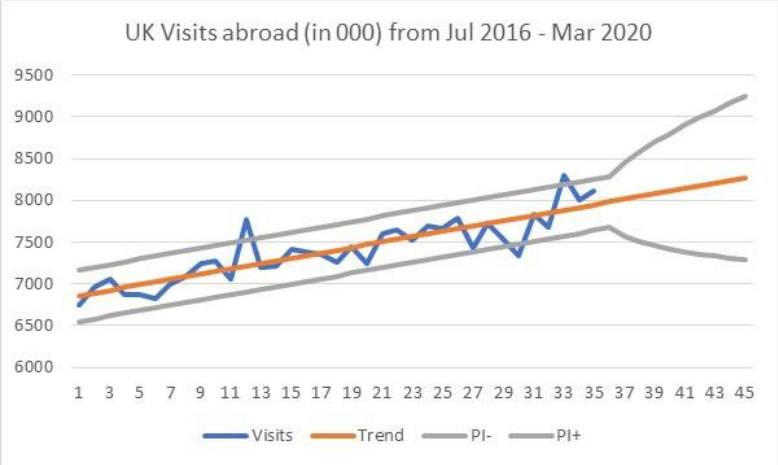

Figure 1

How did we calculate our historical confidence interval, how do we calculate the future prediction interval and why do these formulae work?

To answer these three simple questions, we'll go back to some very basic statistics and refresh what we know about normal distributions. We'll keep it light, with minimum formulae, and maximum visuals and examples.

Basic statistics refresher (just a few points)

You might remember from your basic statistics course that if a dataset (either a sample or the whole population) is distributed in a way that follows a normal distribution, then the mean (for population we use a symbol μ and for a sample we use a symbol $\bar{x}$) is the central value in this distribution. The standard deviation (for population we use a symbol σ and for a sample we use a symbol s) will tell us how individual values are scattered around this mean. If we had a normal distribution (sample data or the whole population), then it will always be that:

a) $\mu \pm 1\sigma$ covers 68.3% of all the values in this dataset ($\bar{x} \pm 1$s for samples)
b) $\mu \pm 2\sigma$ covers 95.4% of all the values in this dataset ($\bar{x} \pm 2$s for samples)
c) $\mu \pm 3\sigma$ covers 99.7% of all the values in this dataset ($\bar{x} \pm 3$s for samples)

We can graphically present this as in Figure 2:

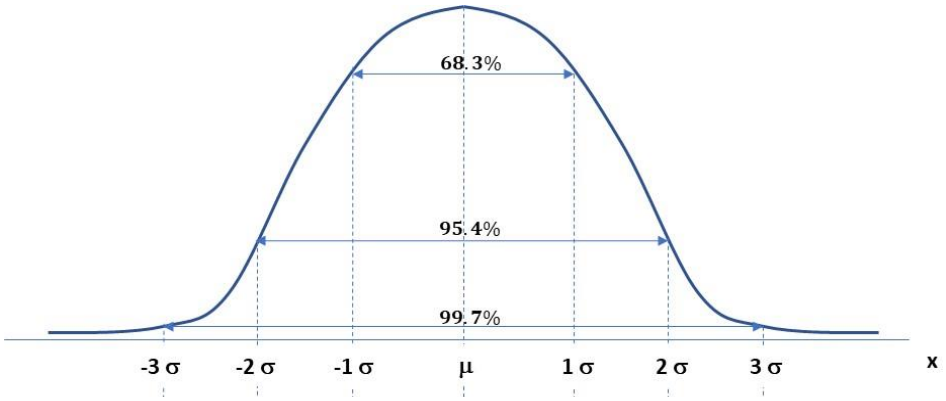

Figure 2

Let's add one point that might sound as a digression, but it isn't. Imagine that you have two datasets, and let's assume that they are both normally distributed. However, one dataset consists of people's weights expressed in kg and the other one of people's heights expressed in cm. How do you compare these two datasets? You have to somehow standardize them. The way to convert any distribution, regardless of the units it uses, into a standardized distribution, is to use the z-values or the so-called z score. The formula for this conversion from x to z is very simple:

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

Example: Let's say we have the mean value $\mu = 20$ and the standard deviation $\sigma = 5$. One of the values in this dataset is x = 22. To convert this x into z, we just plug the values into equation (1).

$$z = \frac{22-20}{5} = \frac{2}{5} = 0.4$$

The value of x=22 converted into a standardized score becomes z=0.4.

What else does this value z tell us? First of all, if we converted our normally distributed dataset into standardized distribution, then the mean value will always be equal to zero. In other words, the mean is expressed as z = 0. The second thing is that the standard deviation is also not expressed in the original units anymore and it is always equal to one. So, z = 1 is the value of the standard deviation of the standardized distribution.

When you convert your dataset into the standardized distribution, it is always true that in this dataset:

a) The range between $\pm$ 1z covers 68.3% of all the values
b) The range between $\pm$ 2z covers 95.4% of all the values
c) The range between $\pm$ 3z covers 99.7% of all the values

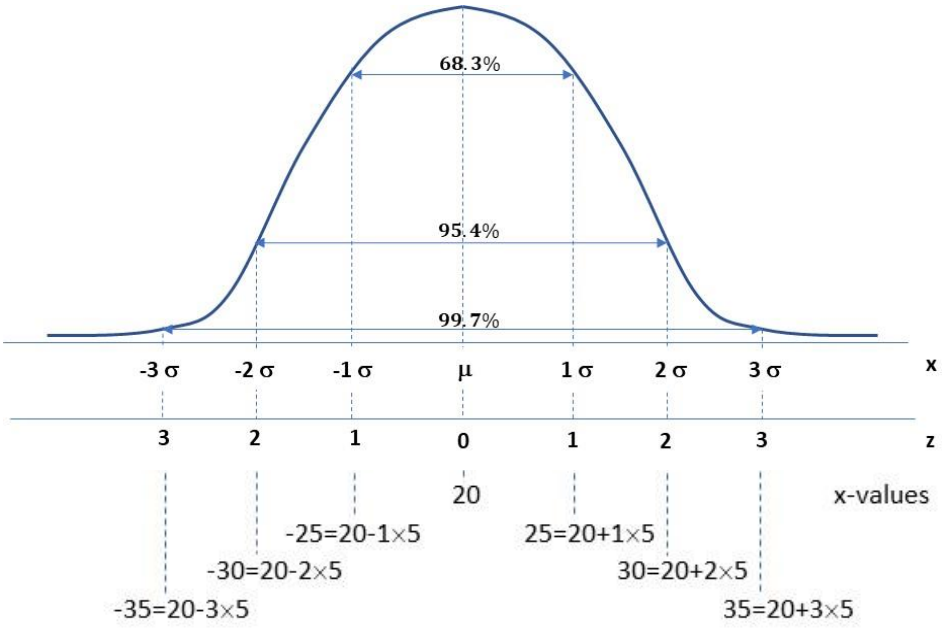

Figure 3

If you do not like these decimal points for percentages and you prefer some round numbers, then you can also say that:

d) The range between $\pm$ 1.64z covers 90% of all the values
e) The range between $\pm$1.96z covers 95% of all the values
f) The range between $\pm$ 2.58z covers 99% of all the values

OK, now we understand the z-score, it is time to return to the main objective of this paper. So far, we assumed that we know the value of μ and σ. However, most of the time you only have a sample, and you calculated the value of $\bar{x}$ and s, but you have no idea what the true value of μ and σ is. Can you use your $\bar{x}$ and s to estimate reliably the true value of μ and σ? Yes, you can. Here is how.

Imagine that you took many, many samples and for every sample you calculated the value of $\bar{x}$. In a way, you now have a dataset consisting of many $\bar{x}$ values, each from a different sample, and these values of many $\bar{x}$ also form a distribution. The mean value of this distribution (called the sampling distribution) of numerous values of $\bar{x}$ will in fact coincide with the true value of the population mean μ. This is great but hold your breath.

If you calculate the standard deviation of such a new dataset (sampling distribution) that consists of numerous $\bar{x}$ from these numerous samples, you will get the value that is equivalent to:

$$SE = \frac{s}{\sqrt{n}} \tag{2}$$

This standard deviation of the distribution of all values of $\bar{x}$ is also called the standard error of the estimate, or as often abbreviated, just **the standard error**.

So, the standard error depends on the sample standard deviation s and the size of the sample n. What immediately jumps from the formula is that the size of the sample n will have a big impact on the standard error. If n is very, very large, then the standard error will be very small. Conversely, small samples will generate a large standard error.

What we have here is the beginning of inferential statistics that says that it is possible to estimate the population parameters from the sample parameters. Our standard error is an estimator of some true value. The good news is that you do not have to take many, many samples to calculate many means $\bar{x}$ and to calculate SE for every sample. A theorem called the **central limit theorem** states that you can do the estimates of the whole population just based on one single sample you took.

What is the central limit theorem? The theorem is very practical in the sense that it states that it does not matter if your sample does not follow a normal distribution. As long as it is large enough, it is expected that the means from many samples (if you took them) would be normally distributed. This implies that everything we said above applies to any kind of sample regardless of the distribution it follows. This is a very good news.

Just one additional point, that also might sound like a diversion, but it is not. If we took many samples and if we look at the distribution of all the values of $\bar{x}$ (sampling distribution), these values of $\bar{x}$ could also be expressed in various units (could be collection of weight averages or collection of height averages). This means that these values of $\bar{x}$ could also be converted into standardized score z. The formula is identical to equation (1) except that the value of x is substituted with $\bar{x}$ and the value of σ is substituted with SE:

$$z = \frac{\bar{x}-\mu}{SE} \qquad (3)$$

Now we finally have a tool to deal with any kind of sample and make reliable estimates of the population parameters on the basis of one single sample. Before we do this, let's just look at these z-values.

To see what percentage under the curve corresponds to different values of z, the tables of all z values under the normal standardized distribution are published. For example, if z = -1, the following can be visualised:

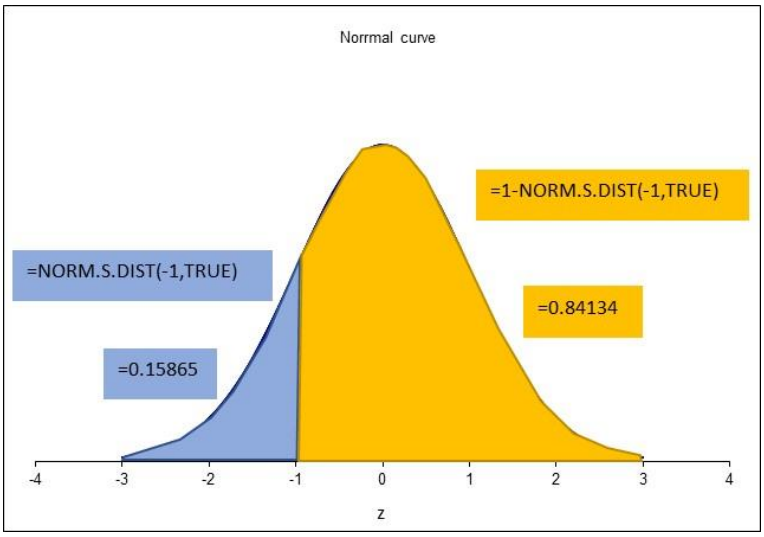

Figure 4

Looking at left of z = -1, we have 15.86% of all the datapoints, and looking to the right we have 84.13% of all the values. The other way to say the same is: there is a 15.86% probability that all the values in this dataset are below z = -1 and there is 84.13% probability that all the values from this dataset are larger than z = -1.

What happens if we subtract these values, i.e. 84.13 - 15.86? We get 68.3. Remember, we stated that 68.3% of all the values should be between $\pm 1\sigma$. Effectively we have the following picture:

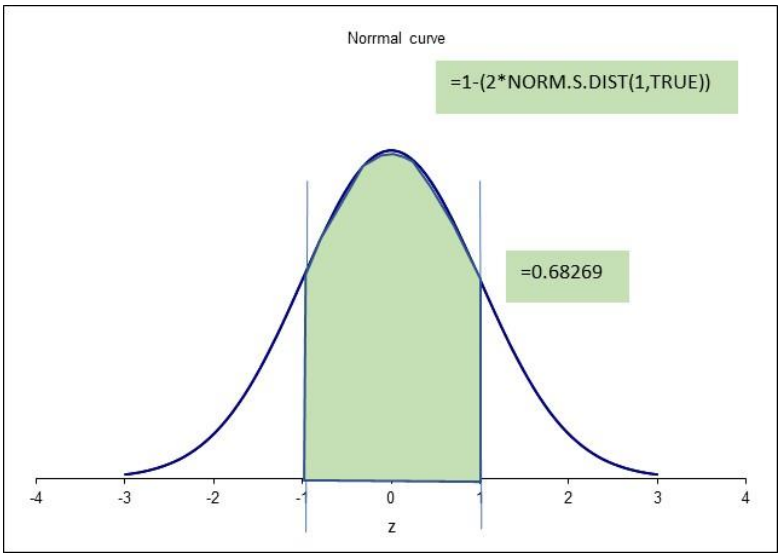

Figure 5

You noticed in Figure 4 and 5 some functions. They are Excel functions to calculate the z-values, which means that we do not even need the tables. To calculate the percentage of all the data to the left of a z-value, we use =NORM.S.DIST(-1,TRUE) and to calculate the percentage to the right of, we use =1-NORM.S.DIST(-1,TRUE). Equally, to calculate the percentage under the curve from two z-values ($\pm$1 in this case), we can use =1-(2*NORM.S.DIST(1,TRUE)). For any other z-value, such as 1.64 for example, just replace 1 in the function =NORM.S.DIST() with this value.

As a side note, if you need to calculate the z-value from the percentage under the curve, then the =NORM.S.INV() function is used. For example, for the level of significance of 0.05 (i.e. 95% confidence level), the z value is calculated as =ABS(NORM.S/INV(0.05/2)), which returns z=1.96. For α=0.1 (i.e. 90% confidence interval), z= ABS(NORM.S/INV(0.1/2))=1.64. And finally, f or α=0.01 (i.e. 99% confidence interval), z= ABS(NORM.S/INV(0.01/2))=2.58.

Just to say it one more time, for every value of z, you can determine what percentage of the curve is covered to the left of this value z. The alternative phrase is: “the value of z will determine the probability that the dataset is below this value of z”. If z-value determines the probability that other datapoint are below and above that z-value, then we can use it for estimation purposes.

Now we have a breakthrough! If we use these simple values of z, we can **estimate** the probability that all other values in this dataset are below this z value.

Interval estimates

Let’s go back to equation (3). As it stands, it is solved for z. What if we rearrange it and solve it for μ? Here it is. First, we move everything in the single line: $z \times SE = \bar{x} - \mu$ and then we express it as a function of μ: $\mu = \bar{x} - z \times SE$.

What does this equation now say to us? It states that the true value of μ can be calculated from the estimated value of $\bar{x}$ and the value of z multiplied by the standard error SE. In fact, this equation

shows just one side of the story. A proper statement would be the expression where z and SE create an interval around $\bar{x}$. In other words:

$$\mu = \bar{x} \pm z\,SE \tag{4}$$

This is an important statement. It says that if you can calculate a parameter from a sample (in this case $\bar{x}$), and if you can calculate the standard error SE (as, $SE = \frac{s}{\sqrt{n}}$), then you can estimate the true value for this parameter that describes the whole population (μ). We can convert this statement into a generic formula:

$$\text{True value} = \text{Estimate} \pm z\,\text{SE} \tag{5}$$

Here is another important point. We have shown above that z defines the percentages under the curve, or the alternative expression, z defines the probability that all the dataset values are below of a particular value of z. This means that for any given value of z, we have in fact defined the probability of our estimates. For z = ±1, the probability will be 68.3%, for z = ± 1.64 the probability will be 90%, for z = ± 1.96 the probability will be 95%, for z = ± 2.58 the probability will be 99%, etc.

Example: We calculated the probabilities for some more commonly used z-values

	I	J	K	L	M	N
1	Value of ±z	Area under the curve				Percentage
2	1	0.683	=1-(2*(1-NORM.S.DIST(I2,TRUE)))			68.3%
3	1.64	0.899				89.9%
4	1.96	0.950				95.0%
5	2	0.954				95.4%
6	2.58	0.990				99.0%
7	3	0.997				99.7%

Figure 6

Now we know that z determines the probability of the estimate and SE determines the width of the interval (don't forget, a large sample will have smaller SE, so the interval is much narrower). This means that we have in fact created a confidence interval that is "backed" by the probability level. This is the reason why we use the phrase "at the confidence level of 95%, the confidence interval is ± xx". An alternative expression is: "we are 95% confident that the true value of μ is in the interval of $\bar{x} \pm z\,SE$, where z = 1.96". Clearly the confidence interval will change, depending on the value of z we select.

Before we proceed to the true objective of this paper, which is the calculation of the prediction interval, just one last digression. From personal experience, this often confuses people that tackle this material for the first time. Let us explain this using an example.

Example 1: Let's assume that the average weight for parcel handled by a Post Office is μ=125 g and the standard deviation is σ=4 g. If we take one package that happens to be 124 g (x = 124), we can convert this into a z-value using equation (1):

$$z = \frac{124-125}{4} = -0.25$$

For z= -0.25, we calculate the probability of p=0.401 (calculated using Excel function =NORM.S.DIST(-0.25,TRUE). This means that 40.1% of all individual packages handled in this Post Office are lighter than 124 g.

Example 2: Now let's say that we took a sample of 25 packages and established the average of 124 g ($\bar{x}$=124). We still know that the average weight for all packages is μ=125 g and the standard deviation is σ=4 g. From this, using equation (2) we can calculate SE, which is 0.8 ($SE = \frac{4}{\sqrt{25}} = 0.8$). If we now apply equation (3) we get the z-value of -1.25:

$$z = \frac{124-125}{0.8} = -1.25$$

For z=-1.25, we calculate the probability of p=0.106. This means that on average, based on sample of 25, 10.6% of all packages are lighter than 124 g. What is going on? This is a different value from the one above. Let's use just one more example.

Example 3: This time we have a sample of 100 packages, again with the average weight of 124 g ($\bar{x}$=124) and the average weight for all packages is still μ=125 g and the standard deviation σ=4 g. We'll use again equation (2) to calculate SE, which is now SE=0.4. If we now apply equation (3) we get the z-value of -2.5. This gives us the p value of p=0.006. The conclusion now is that on average, based on the sample of 100, there are only 0.6% of all packages lighter than 124 g. Can we explain this? Yes, the following paragraph explains it.

In Example 1 we only had one single value of x, which is one single package of 124 g. In Example 2 we had a sample of 25 packages, so our average weight of 124 g for this sample became more representative than the single package weight. In Example 3 we had a sample of 100 packages, so our average of 124 g became even more important as now we are getting more convinced that this is quite representative average.

Essentially the weight of one single package is far less representative that the average weight from a sample of 25 packages. Equally, the average weight for a sample of 25 packages is less representative than the average package for a sample of 100 packages. Think of it as a chance of how far away the weight of one single package is from the true mean of all the packages. With the sample of 25, the chances are that their average will be not so far away to the true average. The average from the sample of 100 will have a chance of being even closer to the true average weight of all parcels for this Post Office. This is exactly what the p values above are telling us. It shows us that the larger the sample, the chances are that the average value will be closer and closer to the actual population average.

Everything we said here is part of basic statistics that deals with descriptive parts and inferential parts of how to estimate the population mean μ from the sample mean $\bar{x}$. Fortunately, this all applies to many other areas, including predictions and forecasting. The only difference is that we substitute the mean values with the prediction values.

Standard errors for predictive models

As we engage in building predictive models and forecasting, we must not forget that the forecasts our model will produce are effectively just estimates of some true future value. In this respect, the forecasts ŷ is not different from $\bar{x}$. While $\bar{x}$ is an estimate of true value μ, ŷ is an estimate of true value y. This implies that our forecasts also require confidence interval. However, the name for such confidence interval is called a prediction interval.

Prediction interval is calculated exactly in the same fashion as the estimates of the true mean:

Future actual value = Estimate of the future value $\pm$ z SE

As before, the value of z will determine the level of confidence. For 95% confidence level we need z = 1.96. The value of SE will determine how wide this prediction interval should be.

Example: Let's say that we produced only three forecasts and that they are: $\hat{y}_1 = 100$, $\hat{y}_2 = 110$ and $\hat{y}_3 = 120$. If SE = 10 and z=1.96, then our prediction interval becomes:

For $\hat{y}_1$ = between 80.4 and 119.6, which is $100 \pm 1.96 \times 10$

For $\hat{y}_2$ = between 90.4 and 129.6, which is $110 \pm 1.96 \times 10$

For $\hat{y}_3$ = between 100.4 and 139.6, which is $120 \pm 1.96 \times 10$

In other words, although our model forecasts for the next period was 100, we can claim with 95% certainty that the actual value might be somewhere between 80.4 and 119.6.

There is only one difficulty with this approach. It assumes that the standard error will stay constant until the end of the forecasting horizon. We intuitively know that the further into the future we try to extrapolate something, the more uncertain the outcome is likely to be. So, the answer is that the prediction interval has to be wider and wider, the further into the future we extrapolate our forecasts. Although the interval gets wider, the confidence level stays the same, which is logical.

If we conduct just simple regression analysis, there is a nice formula that takes care of that. The formula is:

$$SE_{\hat{y},x} = SE_{\hat{y},y}\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \tag{6}$$

The symbol for x refers to the independent variable and $\bar{x}$ is the independent variable mean, whilst $SE_{\hat{y},y}$ is just the standard error for predictions ŷ, as given below:

$$SE_{\hat{y},y} = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n-2}} \tag{7}$$

If used in the context of linear regression, then the function in Excel that is used for $SE_{\hat{y},y}$ is =STEYX(). Otherwise, just a bit more complex formula is needed:

=SQRT(SUMXMY2(array1,array2)/COUNT(array2)),

where array1 are all the x values and array2 all the y values.

In line with equations (4) and (5), we have the Prediction Interval (PI) calculated as:

$$PI = \hat{y} \pm z\, SE_{\hat{y},x} \tag{8}$$

However, for any other forecasting model that does not comply with the least squares method, equation (6) for SE cannot be used. So, what is the solution? Unfortunately, all analytical solutions are too complicated and will depend on the specifics of the forecasting method. Fortunately, there is a rule of thumb, or a workaround that can be applied.

Prediction interval beyond the first forecast

First of all, equation (7) is effectively an equation for calculating the root mean squared error (RMSE), with only one exception that in equation (7) we are using n-2 degrees of freedom and it is

customary to use just n number of degrees of freedom for calculating RMSE. The good news is that for longer series it becomes immaterial if we used n-2 or n.

A rule of thumb we referred to implies that RMSE (or, SE) should not be fixed, but that it should grow with the length of the forecasting horizon. In other words, something like this:

$$RMSE_h = \sqrt{h \times MSE} \tag{9}$$

To switch back to SE, we can rephrase equation (9) into:

$$SE_h = SE\sqrt{h} \tag{10}$$

Where h represents the future time periods h=1, 2, … m, and m is the end of the forecasting horizon.

To be absolutely clear, what we are saying here is that we can calculate SE for the historical data and that this value of SE gets "fixed" at the point of making future forecasts. From that point onwards, the SE is "corrected" by the value of $\sqrt{h}$ for all future forecasts, where h are the simple increments starting with 1 and ending with m, where m is the end of the forecasting horizon. SE_h is now a dynamic value that changes with every new future period.

Now equation (8) can now be modified to account for the increase in uncertainty that the length of the forecasting horizon brings and include equation (10) as:

$$PI = \hat{y}_h \pm z\, SE_h \tag{11}$$

Just as a side note, we kept the use of z-values throughout this paper, but it goes without saying that if we used shorter time series (30 observations or less), we would need to use the t-values.

So, to be practical, we'll use a brief example to demonstrate the use of the prediction interval.

> Example: We took from the ONS site a dataset containing the number of UK visits abroad (in thousands) from July 2016 until May 2019. These 35 observations represent our dataset and our intentions were to forecast the next 10 months until March 2020. We stopped at that point as the ONS dataset stops there and we wanted to see how our forecasts compare with the actual values.
>
> Figure 7 below shows the values in column C, the forecasts using simple Excel =TREND() function in column D and $\pm$ prediction interval in columns F and G (rows 10:30 are hidden).
>
> In Figure 8 we show some calcs we used to execute these forecasts and calculate the prediction interval. We have not covered error measures such as ME, RMSE and MPE, so they are here undescribed, just for the benefit of this analysis.

	A	B	C	D	E	F	G	H	I
1	UK visits abroad in 000								
2	**Date**	**Time t**	**Visits**	**Trend**	**h**	**PI-**	**PI+**	**Actual**	
3	Jul 2016	1	6750	6860.2		6551.0	7169.5		
4	Aug 2016	2	6960	6892.3		6583.0	7201.5		
5	Sep 2016	3	7050	6924.3		6615.0	7233.5		
6	Oct 2016	4	6870	6956.3		6647.0	7265.5		
7	Nov 2016	5	6870	6988.3		6679.1	7297.6		
8	Dec 2016	6	6830	7020.3		6711.1	7329.6		
9	Jan 2017	7	6990	7052.4		6743.1	7361.6		
31	Nov 2018	29	7540	7756.8		7447.5	8066.0		
32	Dec 2018	30	7330	7788.8		7479.6	8098.1		
33	Jan 2019	31	7830	7820.8		7511.6	8130.1		
34	Feb 2019	32	7680	7852.8		7543.6	8162.1		
35	Mar 2019	33	8300	7884.9		7575.6	8194.1		
36	Apr 2019	34	8010	7916.9		7607.6	8226.1		
37	May 2019	35	8120	7948.9		7639.7	8258.2	8120	
38	Jun 2019	36		7980.9	*1*	*7671.7*	*8290.2*	7760	
39	Jul 2019	37		8012.9	*2*	*7575.6*	*8450.3*	7690	
40	Aug 2019	38		8045.0	*3*	*7509.3*	*8580.6*	7780	
41	Sep 2019	39		8077.0	*4*	*7458.5*	*8695.5*	7620	
42	Oct 2019	40		8109.0	*5*	*7417.5*	*8800.5*	7520	
43	Nov 2019	41		8141.0	*6*	*7383.5*	*8898.5*	7440	
44	Dec 2019	42		8173.0	*7*	*7354.8*	*8991.2*	7370	
45	Jan 2020	43		8205.1	*8*	*7330.4*	*9079.8*	7050	
46	Feb 2020	44		8237.1	*9*	*7309.3*	*9164.8*	7090	
47	Mar 2020	45		8269.1	*10*	*7291.2*	*9247.0*	5240	
48				=TREND($C$3:$C$37,$B$3:$B$37,B47)					
49						*=D47-($K$5*$K$6*SQRT(E47))*			
50							=D47+($K$5*$K$6*SQRT(E47))		

Figure 7

	I	J	K	L	M	N	O	P	Q
1									
2									
3		alpha=	0.1						
4		Confidence level=	0.9	=1-K3					
5		z=	1.64	=NORM.S.INV(1-(K3/2))					
6		SE=	188.01	=SQRT(SUMXMY2(C3:C37,D3:D37)/(COUNT(D3:D37)-2))					
7		ME=	0	=(SUM(C3:C37)-SUM(D3:D37))/COUNT(D3:D37)					
8		RMSE=	182.56	=SQRT(SUMXMY2(C3:C37,D3:D37)/COUNT(D3:D37))					
9		MPE=	-0.000586	=SUM(((C3:C37)-(D3:D37))/(C3:C37))/COUNT(D3:D37)					

Figure 8

For a 90% confidence level, our forecasts and the prediction interval are depicted as in Figure 9.

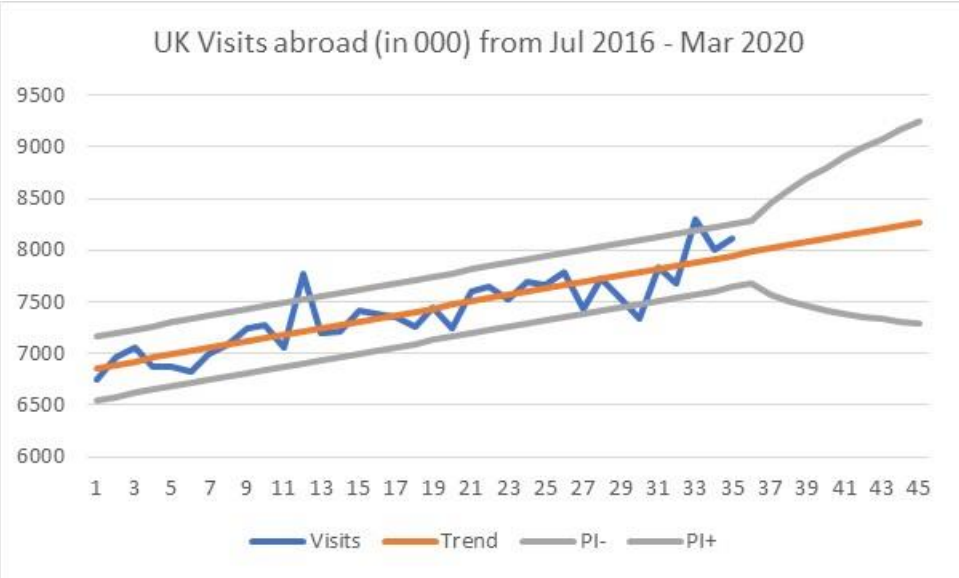

Figure 9

The future prediction interval seems to be surprisingly wide when compared with the historic one. Unfortunately, we know what happened between May 2019 and March 2020. If we include this data into our graph, Figure 10 shows the picture.

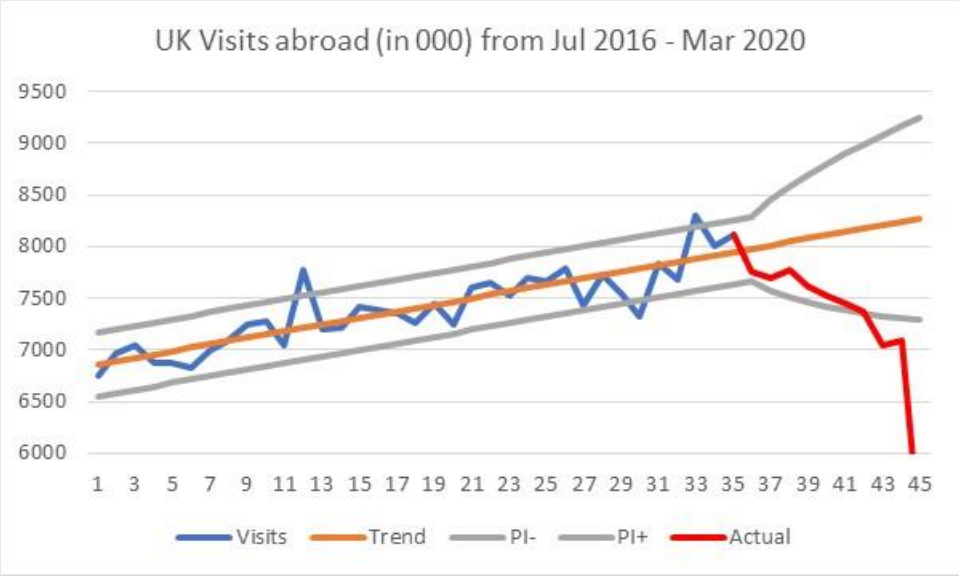

Figure 10

It might be a bit unfair to use this one-in-the-century event such as the pandemic to show how much the longer term future brings uncertainty. Still, our forecasts were in the 90% confidence interval for seven future months before they collapsed. This offers some other lessons about the forecasting horizon, but this could be another paper.

Summary

In this paper, we explained how the values of z and SE are used to define the confidence level and confidence interval in general. Then we applied the same principles to prediction values and extended it to include the growth of uncertainty associated with the length of the forecasting horizon.

Some software packages (including Excel for the moving averages and exponential smoothing) will offer dynamic values of historical SE, which will make the confidence level for the historical values to fluctuate following the fluctuations of the actual values of the dataset. Nothing wrong with that. However, when you reach the end of the dataset and need to calculate the prediction interval, the question is: which value of SE will you use for calculating the future prediction interval? You could use either the last interval value of SE or the overall value of SE. I am not sure I know which one is correct.

Intuitively, the last value of SE should be used if you are dealing with nonstationary data and the overall value of SE if you are dealing with stationary data. However, the jury is still out, and this would require a bit more digging.

In summary, calculating the confidence interval for historical model estimates and prediction interval for the future model estimates (forecasts) is based on some solid descriptive and inferential statistics. The only part where we "dodged" the solid theory was when confronted with an impossible task to include the uncertainty that the future brings into our forecasts. We used a workaround, which not perfect, is still much better than just extending the single value of the SE interval into the future. We created a dynamic value of SE_h that enables us to create an ever-widening prediction interval that corresponds with the intuitive notion that the further into the future you look, the more uncertain it is.

Branko Pecar

Winter, 2021