



ASSOCIATION FOR  
FINANCIAL  
PROFESSIONALS

Underwritten by:



AFP® GUIDE TO

# Leveraging Business Statistics

*Use Cases for Finance*

FP&A GUIDE SERIES



AFP® GUIDE TO

# Leveraging Business Statistics

## *Use Cases for Finance*

### CONTENTS

INTRODUCTION	4
DATA AND METRICS	
Standard Deviation	5
Coefficient of Variation	7
R-squared	8
Correlation Matrix	9
FORECASTING	
Fitting	10
MAPE (Mean Absolute Percentage Error)	10
Naïve Forecast	11
Time-series Analysis	12
ARIMA (Autoregressive Integrated Moving Average)	13
Cluster Analysis	14
Logistic Regression	15
PROBABILITY	
Monte Carlo Simulation	17
TIPS FOR USING STATISTICS WELL	18



Dear AFP Members and Finance Professionals,

Workiva is pleased to partner with AFP to present this guide, “Leveraging Business Statistics, Use Cases for Finance.”

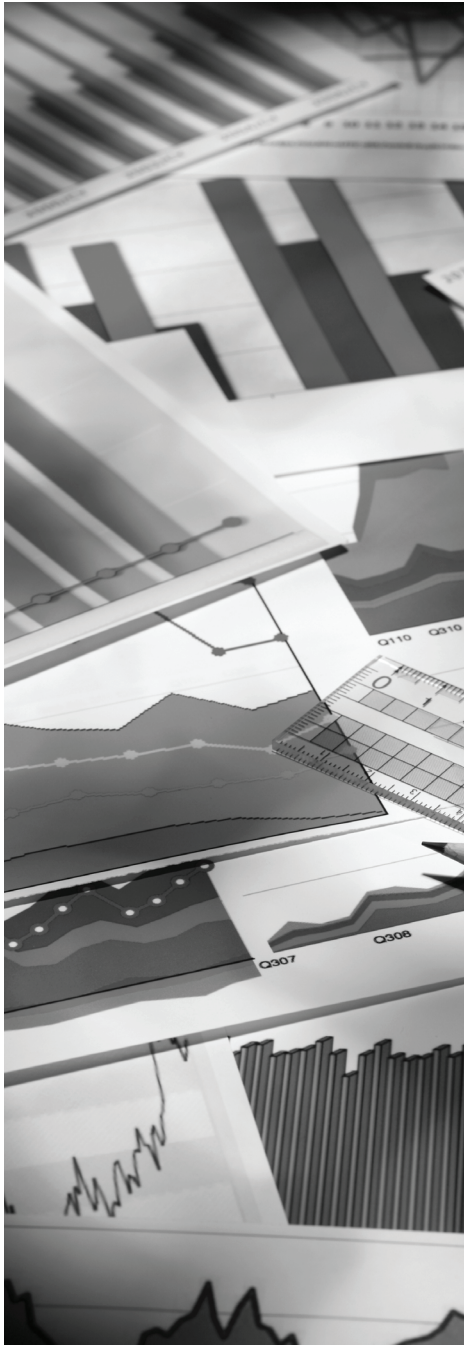
In the following pages, you will learn how FP&A professionals are leveraging a new class of statistical tools to unlock business insights. With an increasing amount of data pouring in from across the enterprise, finance has more information than ever to provide context and predict business outcomes with greater accuracy.

At Workiva, we are committed to helping finance teams access and apply massive amounts data for decision-making with technologies that support data quality, collaboration, and outputs. Finance leaders trust Wdata and Wdesk to enrich and prepare custom datasets for analysis and then present their findings and strategic recommendations in board-ready reports and presentations.

Thanks to modern tools and technologies, today’s office of finance has the opportunity to create a culture driven by data. We hope this guide provides you with practical tools your team can use to turn data into insights that support business growth.

Sincerely,

The Workiva Team



*We live in a world awash with data. Data is proliferating at an astonishing rate—we have more and more data all the time, and much of it was collected in order to improve decisions about some aspect of business, government, or society. If we can't turn that data into better decision making through quantitative analysis, we are both wasting data and probably creating suboptimal performance.*

*—Keeping Up With The Quants, Thomas Davenport & Jinho Kim*

## INTRODUCTION

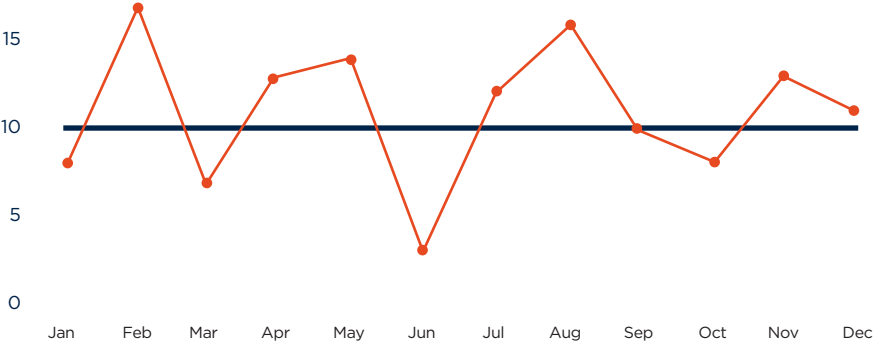
Statistical tools are gaining in importance as the volume, variety, and velocity of data increases. These capabilities are built into spreadsheets, EPM tools, reporting software, and business intelligence packages at rates of increasing sophistication. If we are not using these tools, then we are sub-optimizing. If we are using them incorrectly, then we are in danger of drawing false conclusions.

The rewards for being fluent in statistical tools are many. We become more efficient and effective in everything we do, and can keep pace with the new normal of capabilities driving new operating models. We can remain conversant with our partners in marketing, supply chain, and other areas that are already using statistics in their daily work. This will help us to provide effective challenge to the business and drive decisions as valuable business partners. Without it, we lack a seat at the decision-making table.

This guide is not intended to be a primer or introductory text in leveraging statistical tools. In this guide, underwritten by Workiva, we asked FP&A and finance professionals to identify the statistical tools that are most overlooked or misused; then, we asked them to explain these tools and how they benefit professionals.

## DATA AND METRICS

Good analysis starts with good data, and there are several ways to look at relationships of points to the whole, and for relevance overall. These initial concepts are useful in prioritizing different elements in the data set and separating the outliers. The goal is to avoid GIGO: garbage in, garbage out.

STATISTIC	PRINCIPLE	APPLICATION																										
STANDARD DEVIATION (SD, OR $\sigma$ )																												
<p><b>Definition:</b> Describes the variation or dispersion of data in a set</p> <p><b>Calculation:</b> Square root of the variance, where variance is the difference of each data point from the mean (squared), divided by the number of data points</p> $S = \sqrt{\frac{\sum(X - \bar{X})^2}{N}}$ <p><i>where S = the standard deviation of a sample</i>  <i><math>\Sigma</math> means "sum of,"</i>  <i>X = each value in the data set,</i>  <i><math>\bar{X}</math> = mean of all values in the data set,</i>  <i>N = number of values in the data set.</i></p> <p><b>Use:</b> A low SD indicates a small dispersion of data around the mean; a high SD indicates the opposite</p> <p>If the data are generated according to a normal distribution, 2/3 of all points are within 1 SD above or below the mean, 95 percent within 2 SD, and 99 percent within 3 SD; thus, a data point with an SD=3.5 may be considered an outlier</p>		<p>A Control Chart is used to study a process over time and is particularly useful to continuously improve a process. XYZ Publishing wants to <b>monitor the efficiency of their shipment process</b> and is wondering whether the number of errors they have is acceptable. They decide to track the "days to ship" metric and set up a dashboard to record historical errors and monitor for events that are outside of tolerance, defined by standard deviation. They created the following control chart with boundaries set by the calculated standard deviation.</p> <p><b>Step 1: Plot the metric over a time period</b></p> <p><b>Step 2: Plot the Central Line (<math>\bar{x}</math>)</b></p> <p>The Central Line is the Average (mean) of the selected Measure over a previous period of time indicating how the process has been behaving.</p> <p><b>Avg Days to Ship</b></p> <p>● Days to Ship ● Central Line</p>  <table border="1"> <caption>Avg Days to Ship Data</caption> <thead> <tr> <th>Month</th> <th>Days to Ship</th> </tr> </thead> <tbody> <tr><td>Jan</td><td>8</td></tr> <tr><td>Feb</td><td>17</td></tr> <tr><td>Mar</td><td>7</td></tr> <tr><td>Apr</td><td>13</td></tr> <tr><td>May</td><td>14</td></tr> <tr><td>Jun</td><td>3</td></tr> <tr><td>Jul</td><td>12</td></tr> <tr><td>Aug</td><td>16</td></tr> <tr><td>Sep</td><td>10</td></tr> <tr><td>Oct</td><td>8</td></tr> <tr><td>Nov</td><td>13</td></tr> <tr><td>Dec</td><td>11</td></tr> </tbody> </table>	Month	Days to Ship	Jan	8	Feb	17	Mar	7	Apr	13	May	14	Jun	3	Jul	12	Aug	16	Sep	10	Oct	8	Nov	13	Dec	11
Month	Days to Ship																											
Jan	8																											
Feb	17																											
Mar	7																											
Apr	13																											
May	14																											
Jun	3																											
Jul	12																											
Aug	16																											
Sep	10																											
Oct	8																											
Nov	13																											
Dec	11																											

STATISTIC	PRINCIPLE	APPLICATION
STANDARD DEVIATION (SD, OR $\sigma$ )		

### Step 3: Determine the Standard Deviation ( $\sigma$ )

Similar to the Central Line, we would determine the Standard Deviation over a previous period of time to indicate the variability in the process.

### Step 4: Plot the Upper and Lower Control Limit

The Upper and Lower Control Limits define the boundaries of the process. Ideally, the process should stay within these limits; events outside these bounds are cause for investigation.

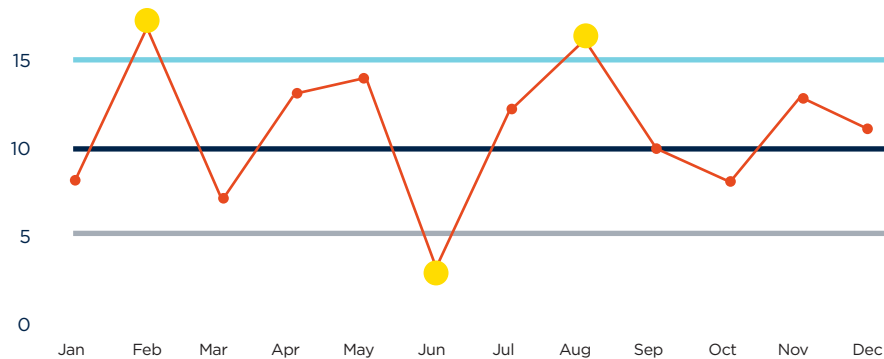
$$\text{Upper Control Limit} = \bar{\chi} + 1\sigma$$

$$\text{Lower Control Limit} = \bar{\chi} - 1\sigma$$

The choice of  $\pm 1\sigma$ ,  $\pm 2\sigma$ ,  $\pm 3\sigma$  or other limits depends on the process and desire to impose control based on acceptable deviation.

#### Avg Days to Ship

● Days to Ship ● Central Line ● Upper Control Limit ● Lower Control Limit



#### Out-of-Control Signals

The most obvious signal that process is not in control is a point outside the control limits. However, other signals and patterns can also point to deficiencies in the process. For example, if the first 20 points fall above the Central Line and next 5 fall below, it would be worth investigating the possible cause for this shift.

—Avi Singh, LearnPowerBI.com

## DATA AND METRICS CONTINUED

STATISTIC	PRINCIPLE	APPLICATION
<b>COEFFICIENT OF VARIATION (CoV)</b>		
<p><b>Definition:</b> Measure of the dispersion of a distribution</p> <p><b>Calculation:</b> Standard deviation/average (mean) of a data set</p> <p>Comparing standard deviation across data sets can be very difficult if the data uses different units or have very different attributes; CoV can standardize this analysis by showing SD relative to its data</p>	<p>An internal business client of FP&amp;A wanted to <b>create a driver-based forecast to predict cost</b> and planned to track 15 metrics of business activity. Pete Geiler, Director of FP&amp;A and member of AFP's FP&amp;A Advisory Council, assembled monthly expenses and operational metrics over a 24-month period. He wanted to use the average (mean) values of the predictors to forecast costs, but the standard deviations varied widely. He then applied the coefficient of variation to determine the correlations relative to the mean; a smaller CoV implies less dispersion of the data around the mean and is therefore a good predictor in percentage terms.</p> <p>Geiler calculated the costs driven by various metrics over the period, then calculated the mean, standard deviation, and coefficient of variation for various other metrics—e.g., membership activities, cost per call, and claims processing. The CoV for membership activities was the lowest among the variables analyzed, showing that changes in these activities had the most significant impact on costs.</p> <p>"The dispersion of the average for this metric was smallest, indicating that a single data point would be a more reliable predictor of costs than other metrics. I needed the coefficient of variation to normalize the data from different metrics with different dispersion characteristics," said Geiler.</p>	

### Healthcare Insurance Example

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	TOTAL				
<b>Metric Volume</b>																	
Contact Minutes	7,000,000	6,500,000	7,200,000	7,100,000	7,300,000	6,900,000	6,800,000	7,300,000	7,000,000	7,300,000	7,600,000	7,700,000	85,700,000				
Membership Transactions	1,500,000	1,350,000	1,600,000	1,500,000	1,600,000	1,500,000	1,600,000	1,600,000	1,500,000	1,600,000	1,700,000	1,700,000	18,750,000				
Claims Processed	900,000	750,000	875,000	850,000	875,000	850,000	825,000	850,000	825,000	850,000	825,000	850,000	10,125,000				
<b>Costs</b>																	
Contacts	\$5,000,000	\$4,900,000	\$5,000,000	\$4,900,000	\$5,000,000	\$4,900,000	\$5,000,000	\$5,000,000	\$4,900,000	\$5,000,000	\$5,000,000	\$5,100,000	\$59,700,000				
Membership	\$3,000,000	\$2,800,000	\$3,200,000	\$3,000,000	\$3,100,000	\$3,000,000	\$3,150,000	\$3,200,000	\$3,000,000	\$3,200,000	\$3,400,000	\$3,400,000	\$37,450,000				
Claims	\$1,500,000	\$1,300,000	\$1,600,000	\$1,500,000	\$1,650,000	\$1,400,000	\$1,700,000	\$1,800,000	\$1,400,000	\$1,500,000	\$1,400,000	\$1,500,000	\$18,250,000				
<b>Cost per Metric</b>																	
															<b>Avg</b>	<b>SD</b>	<b>CoV</b>
Contacts	\$0.71	\$0.75	\$0.69	\$0.69	\$0.68	\$0.71	\$0.74	\$0.68	\$0.70	\$0.68	\$0.66	\$0.66	\$0.70	0.70	0.03	0.040	
Membership	\$2.00	\$2.07	\$2.00	\$2.00	\$1.94	\$2.00	\$1.97	\$2.00	\$2.00	\$2.00	\$2.00	\$2.00	\$2.00	2.00	0.03	0.015	
Claims	\$1.67	\$1.73	\$1.83	\$1.76	\$1.89	\$1.65	\$2.06	\$2.12	\$1.70	\$1.76	\$1.70	\$1.76	\$1.80	1.80	0.15	0.083	

## DATA AND METRICS CONTINUED

STATISTIC	PRINCIPLE	APPLICATION
-----------	-----------	-------------

### R-SQUARED, Also referred to as the coefficient of determination

**Definition:** The strength of correlation between a model (or formula) and the set it is trying to explain

**Calculation:**  $1 - (\text{Variation explained by your model} / \text{Total Variation})$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

**Use:** Expressed between 0 and 1, or 0 percent and 100 percent; a higher number shows your model is a better fit (the greater the amount of variation in the data can be explained by your model)

**Caveat:** Do not dilute your model by adding extraneous variables; additional independent variables will improve the  $R^2$ , even if they are weakly correlated

This model shown (sloped line) has an  $R^2$  of 93 percent because it explains the individual data points very well

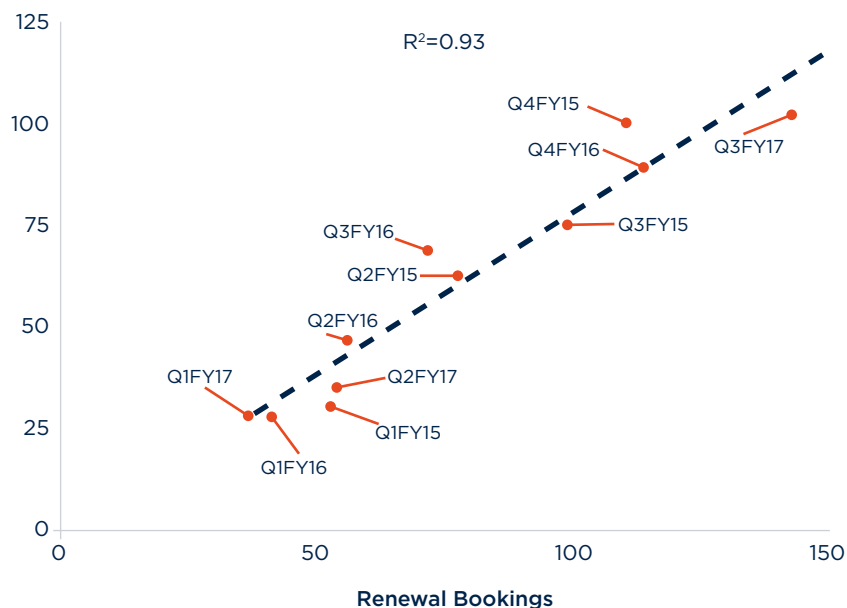
Pankaj Tamrakar, Manager, BMC Software, wanted to improve his sales forecasting, and specifically to analyze **how are new bookings correlated with renewal bookings**. “FP&A always wants to get ahead of the game and predict the future. After a quarter closes, we focus on what happened, and how can we get insight from all this historical data?”

While the immediate challenge was to forecast sales, Tamrakar went “upstream” to look at the operational flow from pipeline opportunities to new bookings. Then, using regression analysis in Tableau, he found a strong correlation, defined by a high  $R^2$ , between new bookings and renewal bookings.

The first insight was that acquiring new business from new customers required generating much larger sales pipeline. Although renewing customers are also the source of new business, they generally require less pipeline and resources. Sales could leverage this insight to allocate resources more efficiently by optimizing the sales coverage ratio for current and future quarters.

The second insight is that the  $R$ -squared described relationship at an equilibrium, but the change to the variables will lead to a new equilibrium. BMC may need to re-evaluate that relationship again in the future after changing the sales coverage.

**New Bookings**





STATISTIC	PRINCIPLE	APPLICATION																								
<b>CORRELATION MATRIX</b>																										
<p><b>Definition:</b> A table that shows correlations among variables in a regression to identify multicollinearity</p> <p><b>Calculation:</b> Calculate the correlation coefficients (<math>r</math>) for all variables</p> $r = \frac{\text{Covariance}(x,y)}{S.D.(x)S.D.(y)}$ <p><b>Use:</b> “Multicollinearity” is a challenge in regression models where the independent “predictors” influence each other, clouding the predictive value, especially volatility, of the overall equation</p> <p>Expressed between 0 and 1—a higher number shows how closely related the variables are to one another, above 0.8 being classified as serious multicollinearity</p> <p><b>Caveat:</b> Do not include independent variables that are duplicates of one another; the model will not be able to distinguish which variables are significant</p>	<p>A real estate firm wanted to rent out space to businesses, and wanted to know how different variables explain an event, such as the selling price of a property. The firm wanted to find which properties are of highest value for commercial use.</p> <p>The firm created a model using indicators of value. However, the model still did not explain which factors were most significant. Next, the firm created a correlation matrix below with the sample data from 150 properties data using Excel/ Data Correlation as part of the Data Analysis add-in from the Analysis ToolPak. The matrix shows a high correlation between property size and rooms, meaning that these two variables may be duplicates contributing the same information in the model.</p> <table border="1"> <thead> <tr> <th>Column 1</th> <th>Price</th> <th>Rooms</th> <th>Property Size</th> <th>Location</th> </tr> </thead> <tbody> <tr> <td>Price</td> <td>1</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Rooms</td> <td>0.531</td> <td>1</td> <td></td> <td></td> </tr> <tr> <td>Property Size</td> <td>0.432</td> <td>0.853</td> <td>1</td> <td></td> </tr> <tr> <td>Location</td> <td>0.543</td> <td>0.744</td> <td>0.577</td> <td>1</td> </tr> </tbody> </table> <p>The first insight was that rooms and property size are duplicates and only one indicator is needed to explain property valuations. Second, when running the estimates with property size and location only, the location becomes the more significant predictor. This leads to the insight that location was the key indicator followed by property size for the valuation of properties.</p> <p>The logic of the high correlation makes intuitive sense. Larger properties have the opportunity for a larger structure with more rooms, and the correlation between them is high.</p> <p>Note that the correlation matrix can also be created “manually” using =COVAR(x,y) and =SDDEV(x) functions in your spreadsheet.</p> <p>—Darren Goonawardana, Acterys</p>	Column 1	Price	Rooms	Property Size	Location	Price	1				Rooms	0.531	1			Property Size	0.432	0.853	1		Location	0.543	0.744	0.577	1
Column 1	Price	Rooms	Property Size	Location																						
Price	1																									
Rooms	0.531	1																								
Property Size	0.432	0.853	1																							
Location	0.543	0.744	0.577	1																						

## FORECASTING

Predictive analytics can be broken down into a simple idea: How can the data of the past inform what will happen next? The explosion of data and computational power has created new ways to create forecasting models. We will begin by explaining the goal of creating a model that “fits” the data, as measured by the MAPE. Then we will dive into some examples of statistical tools that leverage these practices.

STATISTIC	PRINCIPLE	APPLICATION
FITTING		
<b>Definition:</b> “Fitting a model” is creating a relationship between predictors (independent variables) and outcomes (dependent variables)		The XYZ Publishing Company has a series of sales promotions throughout the year; the promotions are essentially the same as the prior year and the business is very stable. XYZ had developed a series of very sophisticated forecasting techniques for each promotion based on promotion parameters and customer behavior, but they wanted to <b>apply the historical data for predictive modeling.</b>
<b>Use:</b> Different models or algorithms may be compared to historical data to see which has the best description (smallest variance) from the data		XYZ purchased a software package that would run more than 20 different analyses—time series, ARIMA, regressions, etc.—to see which one best fit the historical data and compare it to their proprietary model.  The model with smallest historical MAPE (see on page 10) was chosen as the best indicator of future outcomes and used for forecasting purposes, although the test for best fit was conducted periodically in case the data changed and a different curve became more appropriate.
NAÏVE FORECAST		
<b>Definition:</b> Applying the previous period’s forecast to the current forecast without adjustment		With all the new functionality of the software available to the FP&A analysts with the click of the mouse, XYZ became concerned that perhaps they were overthinking their forecasting process. <b>Were their forecast efforts creating value?</b> What if a simple naïve model was better?
<b>Calculation:</b> Forecast period equals prior period actual; variations may introduce seasonality (one year prior) or a lag		If this turned out to be true, it would have had important implications. They should stop expending great effort on their proprietary models or fitting curves and accept a simple, transparent forecast.
<b>Use:</b> This is the “control” group for your forecast; your sophisticated models should do better than the naïve forecast (otherwise you are wasting your time!)		XYZ’s new software included the option of evaluating the fit of the naïve forecast, including one that accounted for seasonality. While it was determined not to be the best fit curve, it also was not the worst as judged by the MAPE calculation. Inside the FP&A team, they used the naïve forecast as a “control” and compared their actual forecasts to this (in addition to actual results) to monitor the accuracy of their predictive modeling.

# FORECASTING CONTINUED

STATISTIC	PRINCIPLE	APPLICATION
-----------	-----------	-------------

## MAPE (Mean Absolute Percentage Error)

**Definition:** The average variance

**Calculation:** The variance of each period, averaged, expressed as a percentage

$$\frac{\sum_{t=1}^n \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right| (100)}{n}$$

**Use:** When comparing the accuracy of various forecasting methods, the one with the lowest MAPE may have the best predictive power

Prior Year	Data		Moving Average				Naïve Forecast				
	Revenue	Actual	Forecast	Variance	% Error	% Error	Forecast	Variance	% Error	% Error	
November	\$1,421,719										
December	\$1,499,453										
January	\$1,937,291	\$1,969,474	\$1,619,487	\$349,987	18%	18%	\$1,937,291	\$32,183	2%	2%	
February	1,704,834	1,839,766	\$1,713,859	\$125,907	7%	7%	\$1,704,834	\$134,932	7%	7%	
March	1,113,434	1,374,198	\$1,585,186	-\$210,988	-15%	15%	\$1,113,434	\$260,764	19%	19%	
April	1,085,156	1,860,024	\$1,301,141	\$558,883	30%	30%	\$1,085,156	\$774,869	42%	42%	
May	1,797,645	1,029,985	\$1,332,078	-\$302,093	-29%	29%	\$1,797,645	-\$767,660	-75%	75%	
June	1,820,403	1,057,060	\$1,567,734	-\$510,674	-48%	48%	\$1,820,403	-\$763,342	-72%	72%	
July	1,080,785	1,599,951	\$1,566,278	\$33,674	2%	2%	\$1,080,785	\$519,166	32%	32%	
August	1,869,714	1,879,559	\$1,590,301	\$289,258	15%	15%	\$1,869,714	\$9,845	1%	1%	
September	1,704,673	1,441,914	\$1,551,724	-\$109,810	-8%	8%	\$1,704,673	-\$262,759	-18%	18%	
October	1,880,891	1,050,455	\$1,818,426	-\$767,971	-73%	73%	\$1,880,891	-\$830,436	-79%	79%	
November	1,971,271	1,425,691	\$1,852,278	-\$426,587	-30%	30%	\$1,971,271	-\$545,580	-38%	38%	
December	1,110,272	1,827,447	\$1,654,145	\$173,302	9%	9%	\$1,110,272	\$717,175	39%	39%	
<b>Sum</b>	<b>\$21,997,540</b>	<b>\$18,355,525</b>	<b>\$19,152,639</b>	<b>-\$797,113</b>	<b>-122%</b>	<b>285%</b>	<b>\$19,076,369</b>	<b>-\$720,844</b>	<b>-140%</b>	<b>424%</b>	
<b>Average</b>	<b>\$1,571,253</b>	<b>\$1,529,627</b>	<b>\$1,596,053</b>	<b>-\$66,426</b>	<b>-10%</b>	<b>24% MAPE</b>	<b>\$1,589,697</b>	<b>-\$60,070</b>	<b>-12%</b>	<b>35% MAPE</b>	

Lower MAPE = Better Fit

STATISTIC PRINCIPLE APPLICATION

TIME SERIES ANALYSIS Simple Moving Average vs. Exponential Smoothing

**Definition:** An attempt to describe a time-based data set over time

**Calculation:** A simple moving average weights all observations equally by taking the average (mean) of the observation:

$$\bar{p}_{SM} = \frac{p_M + p_{M-1} + \dots + p_{M-(n-1)}}{n}$$

An exponential moving average assigns larger weights to more recent observations, so they have more impact on the next (predicted) value

**Use:** Time series models take the current values as inputs and use the past dynamics to forecast the future

“Smoothing” may remove idiosyncratic variation of the time series to make sure it does not have an impact on the forecast

The CFO of a global chemical company wants to improve sales forecasting by finding a best-fit model for historical data. To find sales trends, the CFO compiled the sales data in a spreadsheet, then removed random variation in sales due to extraneous factors, and finally considered these options:

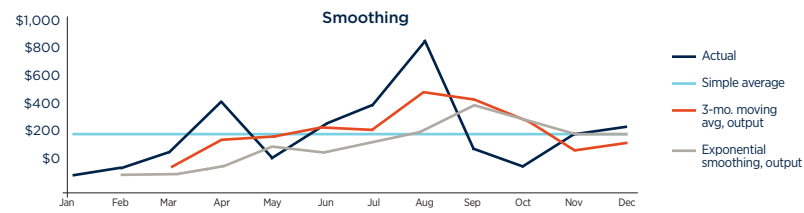
- A simple average of the entire year
- A 3-month moving average that weighted each of the most recent three months equally
- An exponential smoothing option that applies a 70% weight to the previously smoothed month and 30% to the previous data point (applying Excel’s Data Analysis Tool Pak, damping = 0.7)

The first insight was that the sales trend is increasing throughout the year, although masked by the volatility, before declining in Q4. A new sales strategy must be put in place for further growth at year-end.

A second insight is that the moving average has more volatility than the exponential smoothing, which emerged as the basis for the next year’s forecast. The CFO also noted that both smoothing averages can be tweaked and compared for best-fit relative to the data by adjusting the look-back period.

—Darren Goonawardana, Acterys

Period	Actual	Simple average	3-mo. moving average, output	3-mo. moving average, formula	Exponential smoothing, output	Exponential smoothing, formula
1						
2	Jan	\$110	\$345	#N/A	#N/A	#N/A
3	Feb	\$150	\$345	#N/A	#N/A	=C2
4	Mar	\$239	\$345	166	=AVERAGE(B2:B4)	=0.3*B3+0.7*G3
5	Apr	\$540	\$345	310	=AVERAGE(B3:B5)	=0.3*B4+0.7*G4
6	May	\$211	\$345	330	=AVERAGE(B4:B6)	=0.3*B5+0.7*G5
7	Jun	\$390	\$345	380	=AVERAGE(B5:B7)	=0.3*B6+0.7*G6
8	Jul	\$500	\$345	367	=AVERAGE(B6:B8)	=0.3*B7+0.7*G7
9	Aug	\$874	\$345	588	=AVERAGE(B7:B9)	=0.3*B8+0.7*G8
10	Sep	\$253	\$345	542	=AVERAGE(B8:B10)	=0.3*B9+0.7*G9
11	Oct	\$156	\$345	428	=AVERAGE(B9:B11)	=0.3*B10+0.7*G10
12	Nov	\$342	\$345	250	=AVERAGE(B10:B12)	=0.3*B11+0.7*G11
13	Dec	\$380	\$345	293	=AVERAGE(B11:B13)	=0.3*B12+0.7*G12



STATISTIC	PRINCIPLE	APPLICATION
-----------	-----------	-------------

## ARIMA (Autoregressive Integrated Moving Average); Also referred to as the “Box-Jenkins Method”

**Definition:** A predictive time series modeling approach that looks to fit data that’s provided on a periodic basis

**Calculation:** Unlike time series and moving averages, ARIMA uses a regression against its own prior values

**Use:**

**Benefits**

- Does not require advanced statistical knowledge to build
- Can be built using minimal data
- Can be quickly developed, implemented and tested

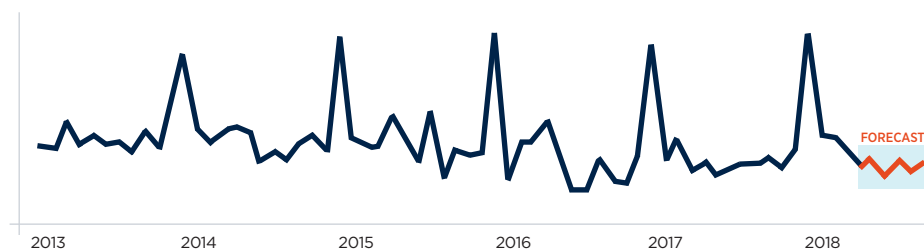
**Challenges**

- Best model fits require stable data over a long time-horizon
- Does not easily provide deep-dive insights into “why” forecasts have their respective fluctuations and trends
- Requires basic understanding of statistics and application that can run ARIMA modeling (e.g., R, Python, SAS Packages)

A mid-market retailer wanted a more **reliable approach to forecasting sales** for the upcoming six months to better understand, predict, and manage the volatility and fluctuations they experience month to month. Cadilus Inc., a firm that provides FP&A services, assembled 10 years of daily sales from the client, restructured the incoming data, and built a modeling capability to predict sales on a 6-month forward-looking basis.

The Cadilus team utilized R packages to complete the project. They kicked off the process by prepping the data for time-series analysis and examining it for significant outliers that could potentially affect model fitting. The team then decomposed the data into the appropriate components: season, trend, cycle [and the residuals]. Once those components were determined, the team ran formal statistical tests to determine *stationarity* (assuring that the series fluctuates in a consistent pattern)—a key requirement for ARIMA modeling. Once stationarity was determined and addressed, the team evaluated the order of parameters (given the components) for the model. Additional evaluations were completed for correlations, autocorrelations and order parameters.

The model was fitted, improved over a few iterations, and the final predictions were made for the upcoming six months. Graphical illustration of forecast is below:



When comparing the monthly actuals to the predicted values, they fell within the prediction range. This allowed the client to feel comfortable that the monthly variations they were experiencing were likely due to seasonal and cyclical components vs. unique events.

—Dan Shin, Cadilus

STATISTIC	PRINCIPLE	APPLICATION
-----------	-----------	-------------

CLUSTER ANALYSIS [K-Means]		
----------------------------	--	--

**Definition:** Using partitioning techniques to group observations together (e.g., customers or products)

**Calculation:** The technique determines the strength or weakness of association between these groups (also known as “clusters”)

The K-Means methodology segments the larger data set around multiple means that define data subsets

**Use:**

*Benefits*

- The outcomes of the analysis are simple and intuitive and can easily be implemented at scale

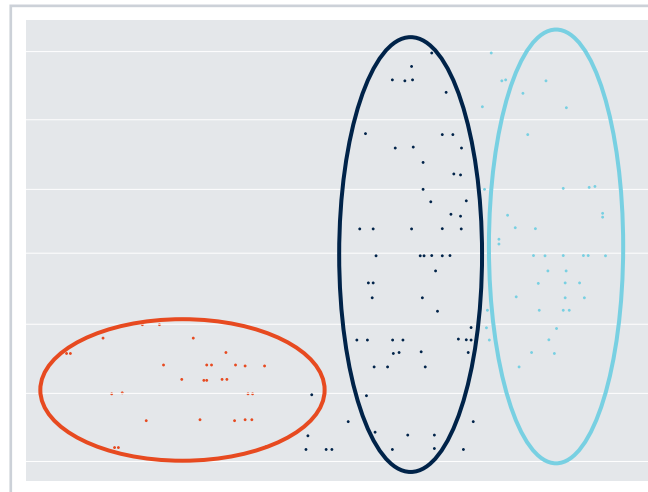
*Challenges*

- Strong associations are determined from the robustness of data variables available
- Requires understanding of statistical tools

A B2B business wanted a more reliable approach to forecasting their future business by **understanding the potential risk and opportunities of customers** currently registered for their platform services versus the current approach of trending and assumption-based ratios. To do so, they wanted to forecast at the customer group level, allowing forecasts to be made with considerations to persistence, attrition, and win-back. Cadilus Inc., a firm that provides FP&A services, collected the full contract history across every customer, both current and lost, to perform the analysis.

The Cadilus team utilized R packages to complete the project. They kicked off the project by organizing and structuring the data for cluster analysis. With the client, the team aligned to three clusters that would represent groups either likely to persist, attrite, or return as a customer.

The K-Means cluster was run on the data and clusters were determined and applied, as show below:



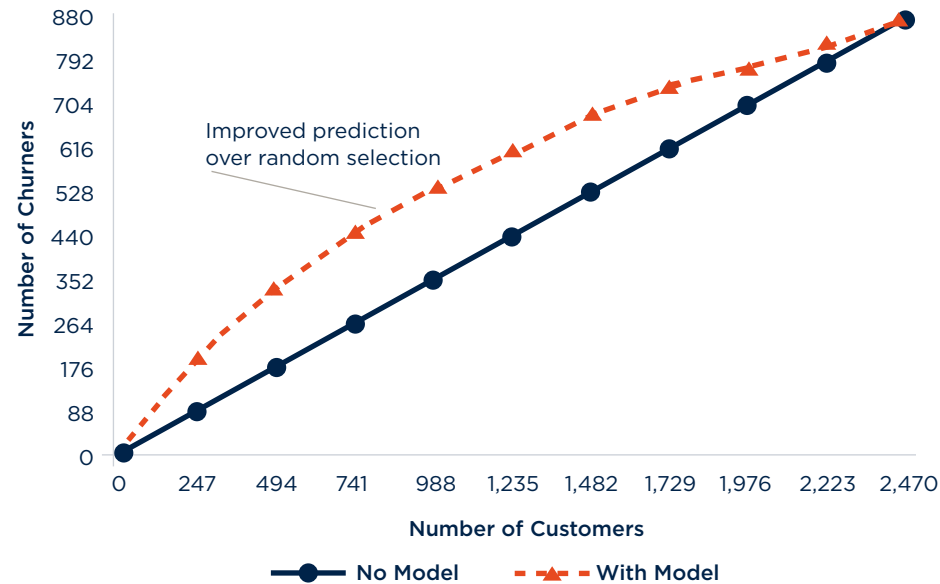
The cluster groupings were applied to the customer data set on an ongoing basis, which allowed the FP&A team to incorporate growth, retention, attribution, and win-back into the monthly sales forecast.

—Dan Shin, Cadilus

STATISTIC PRINCIPLE	APPLICATION																						
LOGISTIC REGRESSION																							
<p><b>Definition:</b> Estimating the relationship of independent variables (predictors) on a dependent (test) variable</p> <p><b>Calculation:</b> A regression describes a dataset by finding a line that minimizes the distance from each datapoint to that line; a logistic regression expresses the probability of the target variable being at one of the two binary outcomes (whereas a linear regression predicts the value of the target variable)</p> <p><b>Use:</b> You can use a logistic regression model to predict binary outcomes, such as whether a team will win or lose, or whether a customer will default on their loan or not</p>	<p><b>Churn rate</b> refers to the percentage of customers ending the business relationship over a given period; the top line of a telecom service company is driven by customer acquisitions and churn (attritions).</p> <p>Jim Gilland, FP&amp;A, is tasked to analyze factors affecting the churn decision, to predict customer churns, and to recommend remedial actions accordingly. Jim gathers data and begins the analysis with three factors: monthly charges (<i>MonthlyBill</i>), whether the account involves any disputes about its monthly charges (<i>Disputed</i>), and whether the account enrolls in paperless billing (<i>Paperless</i>). The decision to either stay with the business or leave represents a binary outcome, so he chooses a logistic or logit model and derives the results below:</p> $\text{logit}(p) = \log \frac{p(\text{churn})}{1 - p(\text{churn})} = -0.752 + 0.001 \times \text{MonthlyBill} + 1.885 * \text{Disputed} - 0.826 * \text{Paperless}$ <p>(A linear regression would be chosen for probabilities between 0 and 1.) Statistical significance tests show that two out of the three variables, <i>Disputed</i> and <i>Paperless</i>, are significant in explaining customer churn. The regression coefficients are the odds of churning for a one unit increase in the predictor variable; technically, the probability of (event occurring) / probability of (event not occurring). To obtain this odds ratio for each coefficient, Jim raises e to the power of the coefficient!</p> <ul style="list-style-type: none"> <li>• The <i>Disputed</i> odds ratio is <math>e^{1.885} = 6.6</math>, i.e., a customer is 6.6 times more probable to churn if the account is <i>Disputed</i> while keeping all other factors constant. [In Excel, the equation is =EXP(1.885)]</li> <li>• The <i>Paperless</i> odds ratio <math>e^{-0.826} = 0.44</math>. When the odds ratio is less than 1, it describes a negative relationship, so we convert to the inverse, or <math>1/.44 = 2.3</math> times, i.e., the customer is 2.3 times more probable to churn if the account is NOT <i>Paperless</i>. [=EXP(-0.826)]</li> <li>• For a customer who has a \$100 monthly bill, a disputed account, and not enrolled in paperless billing, Jim predicts the probability of churning at 78 percent by plugging in these values to solve for <math>p(\text{churn})</math> in the equation.</li> </ul> <p>Building upon a basic understanding of the churning behavior, Jim expands the model to include more relevant factors and applies artificial intelligence through data mining to help retain customers. Specifically, Jim divides the data into training datasets to build models and testing datasets to test and improve the models through machine learning. By grouping existing customers based on their relative probabilities to churn, Jim produces a lift chart, a preferred performance evaluation metric that helps target customers in selected groups. To notch the analysis a step further, Jim uses the churn rate as an input to estimate a customer's lifetime value by a simple customer value multiplier approach. More effective targeting can be designed to prevent profitable customers from churning.</p> <p><small><sup>1</sup>The odds ratio is the ratio of an (event occurring) / (event not occurring), or as stated in the formula, prob (churn) / 1-prob (churn). In log functions, the ratio of log(A)/log(B) can also be expressed as log(A)-log(B). Consider the formula -0.752 + 0.001*<i>MonthlyBill</i> + 1.885*<i>Disputed</i> - 0.826*<i>Paperless</i>. Logistic regressions can only be 0 or 1 (an event happens or it does not), so the ratio is</small></p> <table border="0"> <tr> <td>(2)</td> <td colspan="6">[-0.752 + 0.001*<i>MonthlyBill</i> + 1.885*(1)- 0.826*<i>Paperless</i>]</td> </tr> <tr> <td>(1)</td> <td colspan="6">-[-0.752 + 0.001*<i>MonthlyBill</i> + 1.885*(0)- 0.826*<i>Paperless</i>]</td> </tr> <tr> <td>(2)-(1)=</td> <td>0</td> <td>+</td> <td>0</td> <td>+</td> <td>1.885</td> <td>- 0</td> <td>=1.885, the odds ratio for <i>Disputed</i>.</td> </tr> </table>	(2)	[-0.752 + 0.001* <i>MonthlyBill</i> + 1.885*(1)- 0.826* <i>Paperless</i> ]						(1)	-[-0.752 + 0.001* <i>MonthlyBill</i> + 1.885*(0)- 0.826* <i>Paperless</i> ]						(2)-(1)=	0	+	0	+	1.885	- 0	=1.885, the odds ratio for <i>Disputed</i> .
(2)	[-0.752 + 0.001* <i>MonthlyBill</i> + 1.885*(1)- 0.826* <i>Paperless</i> ]																						
(1)	-[-0.752 + 0.001* <i>MonthlyBill</i> + 1.885*(0)- 0.826* <i>Paperless</i> ]																						
(2)-(1)=	0	+	0	+	1.885	- 0	=1.885, the odds ratio for <i>Disputed</i> .																

STATISTIC	PRINCIPLE	APPLICATION
LOGISTIC REGRESSION		

In conclusion, to reduce the churn rate, Jim recommends 1) more efforts to resolve customer disputes, 2) incentives to increase paperless billing enrollment, and 3) targeting customers identified through a lift chart. The methods Jim employed are applicable to analyzing customer acquisitions, predicting late payment in account receivables, granting loans, investigating employee turnover, even predicting the probability associated with costly equipment failures.



For the above example, this lift chart shows how much more likely we are to reach true churners using the logit model than if we contact customers randomly without a model. We know from an historical training dataset set there are 880 churners out of a population of 2,470 customers. If we contacted 10% (247) randomly without a model, we would expect to reach 10% (88) out of the 880 churners. This is represented by the solid line. Any prediction above this line is an improvement over the “random” approach.

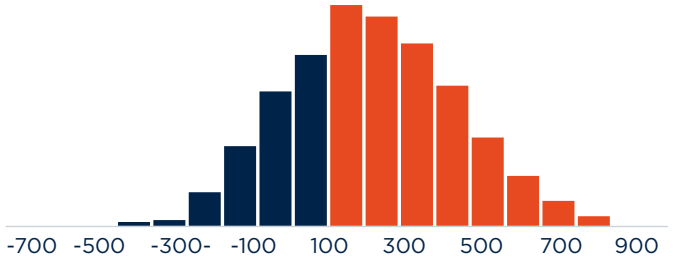
Applying the model, if we rank customers based on their probabilities of churning using the logit model, for the 247 customers in our top decile, we expect to reach 199 churners. This is an improvement over the 88 expected if we used no model, or  $199/88 = 2.27$  x. When more factors are considered, we can potentially lift the reach even further.

—Dr. Xiankui “Bill” Hu, Arkansas State University



## PROBABILITY

The challenge to predictive modeling methodologies is that the future may not look like the past. Alternatively, finance can look to a series of probability related techniques. The most common technique, not represented here, is to create several scenarios and give them weights, such as 15 percent likelihood of an optimistic case, 50 percent of base, and 35 percent downside. Obviously, this applies judgement to forecast, but that is what makes forecasting an art—we mix scientific principles with estimation to investigate something that is unknowable. The next example illustrates that well.

STATISTIC	PRINCIPLE	APPLICATION
MONTE CARLO SIMULATION		
<p><b>Definition:</b> A technique to understand impact of uncertainty by running multiple simulations of an event</p> <p><b>Calculation:</b> A model varies outcomes relative to input variables many times (e.g., 10,000), and produces distributions of possible outcome. Each input has a degree of variability, leading to a distribution of outcomes</p> <p><b>Use:</b> This technique can be applied to diverse problems having probabilistic interpretations, such as revenue guidance, optimistic bias evaluation, derivatives pricing, distribution strategies, stress testing, and so on</p>	<p>A healthcare company has an initial cash balance of \$500,000, expects monthly cash receipts of \$1 million and disbursements of \$900,000. However, both collections and expenditures fluctuate, which creates risk. Jen Lewis, an FP&amp;A professional, is assigned two tasks in preparation for a board meeting this December: <b>1) to project the company's ability to service debt without defaulting, and 2) to help board members better understand risks related to the establishment of a PET/CT facility for its cancer center.</b></p> <p>For the first task of the Monte Carlo Simulation, Jen models monthly cumulative net cash flow (NCF) aggregated from collections and disbursements for the next year. The inputs are monthly collections and disbursements. Discussions with finance teammates reveal collections have a standard deviation of \$250,000, and disbursements have a standard deviation of \$100,000. After fitting historical data with various probability distributions, Jen chooses normal distributions for both inputs and repeats the simulation 10,000 times. The output is set to 1 if the cumulative NCF turns negative during any month and 0 otherwise.</p> <p>Jen finds 1,437 defaults out of the 10,000 runs, implying the probability of default is 14.37 percent, which is high relative to the company's risk tolerance. Jen explores ways to lower the probability of default. Jen repeats the simulation to see the impacts of credit lines (to cover cycles of low collections) on the likelihood of insolvency. She finds that a \$500,000 credit line lowers the default probability to 2.59 percent and a \$2 million credit line virtually eliminates the probability of default.</p> <p>For the second task, Jen models a discounted cash flow (DCF) and calculates the net present value (NPV) of \$121,326. However, this point estimate fails to capture uncertainty or the probability of negative NPV. Jen performs a Monte Carlo simulation again using the DCF model with NPV as the output.</p> <p>Based on the simulation results, Jen plots the distribution of net present value through a histogram. The histogram reveals 31% probability of negative NPV, which is quite high. Jen furthers the analysis using a contribution-to-variance chart that shows how much each input variable contributes to the variation of the project NPV. She finds two demand-related inputs (initial annual number of scans and subsequent growth rates) collectively account for 82.7% of the variation in NPV. She recommends further market research to assess the demand for a PET/CT facility from local and surrounding areas.</p>	<p><b>Distribution of Net Present Value</b> (in thousands \$, as is without adjustment)</p>  <p>—Dr. Xiankui “Bill” Hu, Arkansas State University</p>

## TIPS FOR USING STATISTICS WELL

The power of statistics is the ability to understand the world and make meaning from the increasing piles of data that accumulate around us. The danger is in their misuse, or misunderstanding. Benjamin Disraeli said there are three kinds of lies: “Lies, damned lies, and statistics.” We need to speak this language or be at the mercy of others who do. Here are some tips on how:<sup>1</sup> Questions to ask that will help all your statistics and data projects.

- Which came first, the data or the question?
  - Be careful that you are managing the numbers—don’t let the numbers do the managing for you, or of you.
- What data do you have, and is it the right data for this purpose?
- Can you tell me about the source of the data you used in your analysis?
- Are you sure that the sample data represents the population?
- Are there any outliers in your data distribution? How did they affect the results?
- What assumptions are behind your analysis?
- Are there conditions that would make your assumptions and your model invalid?
- Why did you decide on that particular approach to analysis? Did you use multiple approaches?
- What transformations did you have to do to the data to get your model to fit well?
- Did you consider other approaches to analyzing the data, and if so, why did you reject them?
- How likely do you think it is that the independent variables are actually causing the changes in the dependent variable?
  - Is there additional analysis that can be done?
  - How do you know that the correlation is causality?
  - Did someone check (and duplicate) your findings?

By some estimates, we have created more data in the past three years than we have in all of history, and that is a fraction of what we will create in the next five years. Statistics is a critical tool to work with data, to interpret key messages, make predictions, and take action upon that wealth of data. The capabilities to apply data are expanding as well; it is our responsibility as financial stewards to make sure we are prepared to harness the data and apply our tools appropriately.

<sup>1</sup>This list relies heavily on Keeping Up With the Quants, Davenport & Kim, Harvard Business School Publishing Corporation, 2013. “Good questions about quantitative analyses,” page 171.





## ABOUT THE AUTHOR

**BRYAN LAPIDUS, FP&A**  
Director, FP&A Practice

Bryan Lapidus has more than 20 years of experience in the corporate FP&A and treasury space working at organizations like American Express, Fannie Mae and private equity-owned companies. At AFP he is the staff subject matter expert on FP&A, which includes designing content to meet the needs of the profession and helping keep members current on developing topics. Bryan also manages the FP&A Advisory Council that acts as a voice to align AFP with the needs of the profession.



**ASSOCIATION FOR  
FINANCIAL  
PROFESSIONALS**

## ABOUT THE AFP®

The Association for Financial Professionals (AFP) is the professional society committed to advancing the success of its members and their organizations. AFP established and administers the Certified Treasury Professional and Certified Corporate FP&A Professional credentials, which set standards of excellence in finance. Each year, AFP hosts the largest networking conference worldwide for over 6,500 corporate finance professionals.

4520 East-West Highway, Suite 800  
Bethesda, MD 20814  
T: +1 301.907.2862 | F: +1 301.907.2864

[www.AFPonline.org](http://www.AFPonline.org)



# Critical decisions need reliable data.

Workiva delivers data preparation, management, and reporting solutions that enable you to present the data that drives decisions. Accurate. On time. With context.

When business leaders need the latest numbers, be ready.

Find out how at [workiva.com/afp](https://workiva.com/afp)

**workiva**