Statistical Analysis and Visualization of Customer Visits and Unit Sales

Posted on September 29, 2018

Introduction

This report examines a small dataset linking the number of observed customer visits or purchase occasions with unit sales. The original assignment treated the information as an example of social-science data because purchasing is a form of human behavior shaped by opportunity, need, price, marketing, and the commercial environment. The ten observations show a strong positive linear association: periods with more observed customer activity also have higher unit sales. However, statistical association does not by itself establish that visits are the sole cause of sales. A careful analysis must define the variables, summarize the distribution, fit and interpret the regression correctly, select suitable visualizations, and acknowledge the limits of a very small observational sample (Agresti and Finlay, 2009; Field, 2018; Moore, McCabe, and Craig, 2021).

Dataset and Variable Definition

The first variable is labeled *observed frequency*. In the context of the original explanation, it represents the number of customer visits or purchase occasions recorded during each period. This is the explanatory variable, denoted by x. Unit sales are the response variable, denoted by y. The dataset is preserved below (Agresti and Finlay, 2009).

Observed customer frequency (x)	Unit sales (y)
3	20
6	40
9	55
12	65
15	70

18	89
21	90
24	105
27	110
30	120

The two columns form paired observations. They must not be interpreted as a conventional frequency distribution in which one column lists values and the other states how often each value occurred. Here, each row records one period or observational unit. This distinction is important because it determines which descriptive statistics and graphs are appropriate (Moore, McCabe, and Craig, 2021).

Descriptive Analysis of Unit Sales

Measures of central tendency describe the location of a distribution, while measures of dispersion describe how widely values vary. The mean unit sale is 76.4, calculated by dividing the total of 764 units by ten observations. The median is 79.5, the average of the fifth and sixth ordered values. There is no mode because every sales value occurs once. The mean does not imply that “most” observations equal or closely surround 76.4; it is the arithmetic balance point of the observations (Agresti and Finlay, 2009; Field, 2018).

Statistic for unit sales	Value
Mean	76.4
Standard error of the mean	10.1797
Median	79.5
Mode	None
Sample standard deviation	32.1911

Sample variance	1,036.2667
Excess kurtosis	-0.7380
Sample skewness	-0.3766
Range	100
Minimum	20
Maximum	120
Sum	764
Count	10

The standard deviation of about 32.19 units indicates substantial variation around the mean. Because the values range from 20 to 120, that variability is unsurprising. The skewness is negative, not positive. A value of -0.3766 indicates mild left skew: the lower observations create a somewhat longer lower tail, although the distribution is not strongly asymmetric. Negative excess kurtosis indicates a flatter distribution than a normal distribution in this small sample. Neither statistic should be overinterpreted because ten observations provide limited evidence about the underlying population (Field, 2018; Moore, McCabe, and Craig, 2021).

Scatterplot as the Primary Visualization

A scatterplot is the most informative graph for these paired quantitative variables. Observed customer frequency belongs on the horizontal axis and unit sales on the vertical axis. Each point represents one period. The points rise from the lower left to the upper right and remain close to a straight line. This visual pattern supports a strong positive linear relationship. It also allows the analyst to look for curvature, isolated points, changes in variability, or clusters that a single correlation coefficient could hide (Agresti and Finlay, 2009; Moore, McCabe, and Craig, 2021).

No observation is an obvious extreme outlier relative to the fitted line. That does not mean every point “on” the line is correct and every point away from it is an outlier. A residual is simply the difference between the observed sales value and the value predicted by the fitted model. An outlier requires contextual and statistical investigation rather than a visual label alone (Field, 2018).

Linear Regression Model

Ordinary least squares produces the following fitted equation:

Predicted unit sales = 18.6 + 3.5030 × observed customer frequency

The slope of approximately 3.503 means that within the observed range, an increase of one unit in customer frequency is associated with an average increase of about 3.5 unit sales. Because frequency rises in increments of three in the dataset, a three-unit increase is associated with roughly 10.51 additional sales. This is an average model-based association, not a guarantee for every period (Agresti and Finlay, 2009; Field, 2018).

The intercept is 18.6 units. It is the model's predicted sales value when observed frequency equals zero. The dataset contains no zero-frequency observation, so the intercept is an extrapolated mathematical anchor rather than direct evidence that the business would sell exactly 18.6 units without customers. It also does not measure "all other factors." Those factors are absorbed into unexplained variation and cannot be identified from this simple model (Moore, McCabe, and Craig, 2021).

Regression statistic	Value
Multiple R	0.988405
R square	0.976945
Adjusted R square	0.974063
Residual standard error	5.184330
Observations	10

The correlation coefficient, $r = 0.9884$, indicates a very strong positive linear association. The coefficient of determination, $R^2 = 0.9769$, means that about 97.7 percent of the variation in observed unit sales is accounted for by the fitted linear relationship with customer frequency in this sample. It does not mean that frequency "causes" 97.7 percent of sales or that the model will explain the same proportion in another store or period (Agresti and Finlay, 2009; Field, 2018).

ANOVA and Statistical Significance

Source	df	SS	MS	F	Significance F
Regression	1	9,111.3818	9,111.3818	338.9995	7.797×10^{-8}
Residual	8	215.0182	26.8773		
Total	9	9,326.4000			

The very small p-value provides strong evidence against a population slope of zero under the assumptions of the linear model. Statistical significance must still be separated from practical significance and causal inference. The observations may have been recorded in a sequence, which could introduce time trends or autocorrelation. Promotions, prices, holidays, stock availability, weather, staffing, and the distinction between visitors and purchasers could influence both variables (Field, 2018; Moore, McCabe, and Craig, 2021).

Prediction at a Frequency of 35

Substituting 35 into the fitted equation gives a prediction of approximately 141.21 unit sales. The original estimate of about 140 was therefore close. Nevertheless, 35 lies beyond the observed maximum of 30, so this is extrapolation. A prediction outside the sampled range is less reliable because the relationship may flatten when capacity, inventory, or customer conversion limits are reached. A responsible report should provide a prediction interval, not only a point estimate, and should avoid claiming that the prediction is certainly correct merely because the intercept is constant (Agresti and Finlay, 2009; Moore, McCabe, and Craig, 2021).

Histogram and Distribution

A histogram can summarize the distribution of unit sales, but its bars must contain counts of observations within non-overlapping intervals. The original proposed histogram table repeated increasing customer-frequency values and therefore did not represent valid bin counts. One reasonable corrected grouping is shown below (Field, 2018).

Unit-sales interval	Number of observations

20–39	1
40–59	2
60–79	2
80–99	2
100–120	3

With only ten observations, the shape changes easily when bin boundaries change. A dot plot would often communicate this small dataset more honestly. The negative skewness statistic also contradicts the original description of right skew (Field, 2018).

Why a Pie Chart Is Not Ideal

A pie chart is designed for mutually exclusive parts of a meaningful whole, such as categories of a fixed budget. The ten unit-sales observations are measurements across periods, not categories dividing one total. Although 120 is approximately 15.7 percent of the numerical sum and 20 is approximately 2.6 percent, those percentages do not answer a useful business question. A scatterplot, line chart across time, dot plot, or properly constructed histogram is preferable (Moore, McCabe, and Craig, 2021).

Social-Science Interpretation and Limitations

The data suggest that customer traffic and sales move together, but human purchasing behavior cannot be reduced to one predictor. The observed frequency may itself be influenced by advertising, store location, income, season, product quality, social influence, and availability. Unit sales may also exceed the number of visits because a customer can buy multiple units. The analysis lacks information about how the sample was chosen, whether each period had equal duration, and whether repeated measurements came from comparable conditions (Agresti and Finlay, 2009).

A stronger study would define a unit of observation, gather a larger random or representative sample, record price and promotional variables, distinguish visitors from purchasers, and examine conversion rate and average units per transaction. Residual plots should be reviewed for linearity and constant variance. If data are ordered in time, time-series structure should be assessed. These additions would turn an attractive descriptive relationship into a more defensible decision model (Field, 2018; Moore, McCabe, and Craig, 2021).

Confidence, Prediction, and Replication

A fitted line is only one estimate from one sample. If new periods were observed, the slope and intercept would change. Confidence intervals describe uncertainty about the average relationship, while prediction intervals are wider because they also include period-to-period variation. The original spreadsheet reports a small standard error for the slope, but a practical business forecast should still incorporate inventory constraints, capacity, changing prices, and seasonal demand (Agresti and Finlay, 2009; Field, 2018).

Replication is especially important because the ten frequencies form a perfectly regular sequence from 3 to 30. That design may have been constructed for an exercise rather than randomly observed in normal operations. A second dataset collected prospectively should record the same definitions and duration for every period. The analyst could then test whether the relationship remains linear, whether the slope is stable, and whether a model containing price or promotion improves prediction without overfitting (Moore, McCabe, and Craig, 2021).

Ethical Use of Customer Data

If customer visits are measured through cameras, mobile devices, loyalty programs, or online tracking, the business should consider notice, consent, data minimization, security, retention, and unequal effects. Statistical usefulness does not automatically justify intrusive collection. Aggregated counts may answer the operational question without retaining identifiable movement or purchase histories.

Conclusion

The dataset contains a strong positive linear association between observed customer frequency and unit sales. Unit sales average 76.4, with a median of 79.5 and a standard deviation of about 32.19. The fitted model is $\hat{y} = 18.6 + 3.503x$, with R^2 of approximately 0.977. A frequency of 35 produces a point prediction of about 141 sales, but that estimate is an extrapolation. The corrected analysis rejects the original claims of right skew, causal certainty, and guaranteed prediction. It also identifies the scatterplot—not a pie chart—as the most useful visualization. The results are promising, but the small observational sample requires cautious interpretation (Agresti and Finlay, 2009; Field, 2018; Moore, McCabe, and Craig, 2021).

References

Agresti, A., and Barbara Finlay. *Statistical Methods for the Social Sciences*. 4th ed., Pearson, 2009.

Field, Andy. *Discovering Statistics Using IBM SPSS Statistics*. 5th ed., Sage, 2018.

Moore, David S., George P. McCabe, and Bruce A. Craig. *Introduction to the Practice of Statistics*. 10th ed., W. H. Freeman, 2021.