Statistical Analysis and Visualization of Customer Visits and Unit Sales

Posted on September 29, 2018

View this essay online:
https://academic-master.com/statistical-analysis-and-visualization-of-customer-visits-and-unit-sales/

Introduction

Statistical analysis can turn a small set of observations into a clearer description of how two variables move together, but the quality of the conclusion depends on correct definitions and cautious interpretation. In this dataset, observed customer frequency is paired with unit sales across ten periods, creating two quantitative variables that can be examined through descriptive statistics, a scatterplot, and simple linear regression. The values show a strong positive association: periods with greater customer activity also tend to record more unit sales. That relationship is useful for planning, yet it should not be treated as proof that customer visits alone cause sales because price, promotion, stock availability, seasonality, and customer purchasing behavior may affect both variables. The analysis therefore has four purposes: to summarize the sales distribution, visualize the paired observations, estimate the linear relationship, and explain the limits of prediction from a small observational sample. Proper statistical reasoning distinguishes association from causation, sample results from population conclusions, and mathematically convenient estimates from operationally reliable forecasts (Agresti & Finlay, 2009; Field, 2018; Moore et al., 2021).

Dataset and Descriptive Statistics

Each row represents one observational period rather than a frequency category, so the two columns must be interpreted as paired measurements. The explanatory variable, x, is observed customer frequency, and the response variable, y, is unit sales. The sales values total 764 units, giving a mean of 76.4, while the median is 79.5 because the middle two ordered observations are 70 and 89. No mode exists because every sales value occurs once. The sample standard deviation is approximately 32.19 units, showing substantial dispersion around the mean, and the range is 100, from 20 to 120. Sample skewness is about -0.377, indicating mild left skew rather than right skew, while negative excess kurtosis suggests a flatter distribution than the normal curve. With only ten observations, however, shape statistics are unstable and should not be overinterpreted. The tables below preserve the underlying dataset and descriptive results so that the calculations remain transparent and reproducible rather than being reduced to unsupported narrative statements (Field, 2018; Moore et al., 2021).

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

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

Scatterplot and Linear Relationship

A scatterplot is the most informative primary visualization because it preserves the pairing between customer frequency and unit sales. Placing observed customer frequency on the horizontal axis and unit sales on the vertical axis produces points that rise strongly from lower left to upper right and remain close to a straight line. Ordinary least squares estimates the relationship as predicted unit sales = 18.6 + 3.5030 × observed customer frequency. The slope means that, within the observed range, an increase of one unit in customer frequency is associated with an average increase of about 3.5 sales; because x rises by three units between listed observations, that corresponds to roughly 10.51 additional units sold for each three-unit increase. The intercept of 18.6 is the fitted value when frequency equals zero, but zero is outside the dataset, so the intercept is a mathematical anchor rather than evidence that exactly 18.6 units would be sold without observed customers. Regression coefficients describe the sample relationship and should not be interpreted as hidden measures of every unobserved business factor (Agresti & Finlay, 2009; Field, 2018).

Regression statistic	Value
Multiple R	0.988405

R square	0.976945
Adjusted R square	0.974063
Residual standard error	5.184330
Observations	10

Model Significance and Prediction

The correlation coefficient of approximately 0.988 indicates a very strong positive linear association, while R^2 of approximately 0.977 means that the fitted line accounts for about 97.7 percent of the observed variation in unit sales within this sample. This value does not mean customer frequency causes 97.7 percent of sales, because regression fit is not a causal decomposition. The ANOVA table produces an F statistic of about 339 and a very small significance value, providing strong evidence that the population slope would not be zero if the model assumptions and sampling process were appropriate. Substituting x = 35 into the fitted equation gives a predicted sales value of approximately 141.21. That estimate is useful for illustration but is an extrapolation because 35 exceeds the observed maximum of 30. Business capacity, inventory, conversion rates, or nonlinear demand could change beyond the sampled range. A responsible forecast would therefore report uncertainty and avoid treating the point prediction as guaranteed performance merely because the fitted line is statistically strong (Field, 2018; Moore et al., 2021).

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			

Choosing Appropriate Visualizations

The dataset supports several useful graphs, but each answers a different question. A scatterplot should be used to evaluate the relationship between customer frequency and sales because it preserves paired observations and makes linearity, unusual points, and changing spread visible. A

histogram can summarize the sales distribution, although ten observations are too few to make its shape stable; changing the bin boundaries can alter the appearance substantially. A dot plot would often communicate the small sample more honestly. A pie chart is not appropriate because the sales observations are measurements across periods, not mutually exclusive categories that partition one meaningful whole. The corrected histogram below uses non-overlapping intervals and counts how many sales observations fall into each interval. If the observations were collected over time, a line chart could also be useful for detecting seasonality or trends that simple regression ignores. Visualization should therefore be selected according to the analytical question rather than because software makes a particular chart easy to create. Good graphs clarify structure instead of turning numerical totals into visually impressive but conceptually misleading percentages (Moore et al., 2021).

Unit-sales interval	Number of observations
20–39	1
40–59	2
60–79	2
80–99	2
100–120	3

Interpretation, Limitations, and Better Study Design

The strong fitted relationship is promising, but the dataset is too small and underspecified to support broad business conclusions. It is unclear how the observational periods were selected, whether they were equal in duration, whether customer frequency means visitors or purchase occasions, and whether price, promotions, stockouts, holidays, staffing, weather, or local events changed across observations. A stronger study would define the unit of observation precisely, collect a larger representative sample, record conversion rate and units per transaction, and include relevant predictors such as price and promotional status. Residual plots should be examined for linearity and constant variance, and time-ordered observations should be checked for autocorrelation. Replication is especially important because the x values form a perfectly regular sequence from 3 to 30, which suggests the dataset may have been constructed for an exercise rather than sampled from normal

operations. If customer visits are measured through cameras, loyalty programs, or mobile tracking, data minimization and privacy should also be considered so that useful operational analysis does not depend unnecessarily on identifiable customer histories.

Conclusion

The ten observations show a strong positive relationship between customer frequency and unit sales, with average sales of 76.4 units, a median of 79.5, and a sample standard deviation of approximately 32.19. The fitted regression equation is $\hat{y} = 18.6 + 3.503x$, and R^2 is approximately 0.977, indicating that the line fits this small sample closely. A customer-frequency value of 35 produces a point prediction of roughly 141 sales, but that estimate extends beyond the observed range and should be treated cautiously. The statistical evidence therefore supports a descriptive conclusion that sales rise with observed customer activity, not a causal claim that frequency alone determines sales. Scatterplots and correctly constructed distribution plots communicate the data more appropriately than a pie chart, while the preserved tables make the calculations auditable. The most important next step is not to celebrate the high R^2 but to collect a larger, well-defined dataset containing operational variables that can explain why customer traffic converts into sales and whether the relationship remains stable under different business conditions.

References

Agresti, A., & Finlay, B. (2009). *Statistical Methods for the Social Sciences* (4th ed.). Pearson.

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). SAGE.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2021). *Introduction to the Practice of Statistics* (10th ed.). W. H. Freeman.

Copyright © 2026 Academic Master.