Universal Library of Business and Economics **Research Article**

ISSN: 3064-9951 | Volume 2, Issue 2

Open Access | PP: 01-07

DOI: https://doi.org/10.70315/uloap.ulbec.2025.0202001

Application of Significance Levels for Decision Making in Financial Planning

Angelina Shyltsyna

Financial Services Representative, Barnum Financial Group.

Abstract

This study aims to systematize and substantiate the application of significance levels α in the financial decision-making process and develop a unified approach for adapting traditional statistical criteria to the modern characteristics of economic data. This work synthesizes the historical progression of formalized hypothesis testing—from Pearson's χ^2-test and Fisher's p = 0.05 threshold to contemporary adjustments for heavy-tailed distributions and autocorrelated series. The relevance of this research is driven by the dramatic increase in volumes of high-frequency and nonlinear financial data, which undermines classical asymptotic assumptions. Using examples of the tail dependencies in the S&P 500 and the autocorrelation of weekly NYSE returns, we demonstrate that, without accounting for nonstandard distributional forms and series memory, the probability of false inferences substantially exceeds the declared risk levels. The article details methods for multiple comparison corrections, bootstrap and Monte Carlo simulations, and practical VaR back-testing schemes according to Basel's "traffic-light" methodology. The novelty of this work lies in the comprehensive integration of classical and modern statistical techniques: in addition to conventional p-value thresholds, we propose adaptive boundaries for heavy tails, incorporate adjustments for autocorrelation and heteroskedasticity, and integrate Bayesian and Holm corrections. Practical case studies in event-driven M&A analyses and marketing A/B tests illustrate how adapting the α level enhances strategy reliability and prevents the proliferation of false positives. Our principal conclusion is that rigorous and well-justified application of the significance level α, taking into account the data structure and the nature of the hypotheses under test, transforms the statistical test from a formal procedure into an effective tool for financial planning. Correct selection between one-tailed and two-tailed criteria, adjustment for autocorrelation, application of multiple-test corrections, and simulation techniques enable a balanced trade-off between Type I and Type II errors, thereby minimizing operational costs and strengthening confidence in analytical outcomes. This article will be of value to financial analysts, risk managers, and developers of algorithmic strategies.

Keywords: *Significance Level, P-Value, Hypothesis Testing, Financial Planning, Risk Management, Bootstrap, Monte Carlo, Heavy Tails.*

INTRODUCTION

The notion of the significance level α, which has become the cornerstone of statistical inference and quantitatively justified financial decisions, traces its roots to the early decades of the twentieth century. Karl Pearson introduced the χ^2 criterion for testing goodness-of-fit, laying the foundation for formalized hypothesis testing; thereafter, Ronald Fisher proposed the practical threshold p = 0.05, "convenient as a boundary separating random fluctuations from significant deviations" [1]. The pioneering duality of the null and alternative hypotheses developed by Jerzy Neyman and Egon Pearson supplemented this concept with explicit control of Type I and Type II errors, thus transforming the selection of α into an instrument of rational risk: it defines the probability of making an incorrect investment decision if the market model is misspecified.

The transfer of statistical methodology into finance began in the 1950s, when Harry Markowitz demonstrated that a portfolio's mean–variance characteristics could be optimized systematically; it was at this point that the parametric estimation of variance, and hence confidence in the α level, acquired direct monetary significance. Subsequently, hypothesis testing became integral to empirical market-efficiency studies: event tests for nonzero excess returns

Citation: Angelina Shyltsyna, "Application of Significance Levels for Decision Making in Financial Planning", Universal Library of Business and Economics, 2025; 2(2): 01-07. DOI: https://doi.org/10.70315/uloap.ulbec.2025.0202001.

and regression-based criteria for the Capital Asset Pricing Model (CAPM) relied on fixed significance boundaries when selecting assets and constructing strategies.

During the 1980s and 1990s, the rapid advancement of computing power fundamentally transformed the nature of financial data: intraday quotations expanded by orders of magnitude, and bootstrap and Monte Carlo procedures enabled the estimation of test-statistic distributions without stringent asymptotic assumptions. On these data sets, Gopikrishnan et al. uncovered a power-law form in the tails of S&P 500 returns, with an exponent $\alpha \approx 3$ stable for intervals up to four days [2]. This departure from normality prompted a revision of critical thresholds: minimizing investment errors required adapting α to heavier tails than those implied by the classical Gaussian approximation.

The empirical dependence between successive returns further exacerbated the bias in p-values. Lo and MacKinlay, examining weekly portfolio returns on the NYSE for 1962–1985, identified positive autocorrelation, refuting the random-walk hypothesis and demonstrating that ignoring memory leads to underestimation of actual risk [3]. Financial planning implies adjusting standard criteria (e.g., strategy effectiveness tests) for the dependent data structure; otherwise, the probability of falsely detecting a "persistent" anomaly will exceed the stated significance level.

MATERIALS AND METHODOLOGY

This investigation into the application of significance level α in financial planning draws upon more than twenty key sources, encompassing historical, theoretical, and applied aspects of statistical inference in finance. Foundational starting points include Pearson's and Fisher's works on the χ^2 test, the $p = 0.05$ threshold [1], and the Neyman–Pearson framework for controlling Type I and Type II errors. Empirical characteristics of heavy return tails and the need to adapt significance boundaries are illustrated by the studies of Gopikrishnan et al. on the S&P 500 [2] and Lo and MacKinlay on NYSE weekly returns autocorrelation [3]. The issue of an avalanche of false discoveries under multiple testing is addressed by Harvey, Liu, and Zhu [4]. At the same time, common misinterpretations of p-values are documented by Badenes-Ribera et al. [5]. To understand the impact of α on real-world strategies, we reference reviews on fund performance and the concept of "alpha" (Kumar and Neha [6], Chen [7]), marketing A/B test examples (Reitsnik [8], Ackerson [22]), event-study methodologies for M&A (Bîlteanu [9], Müller [21]), and regulatory frameworks for VaR back-testing (BCBS [10], Accounting Insights [19]). Bootstrap and Monte Carlo procedures for empirical replication of critical thresholds without strict assumptions are detailed by Gopikrishnan et al. [2] and Huang et al. [11], while practical portfolio diversification guidelines derive from Bouslama and Ouda [12]. Finally, the toolkit for computation and visualization includes the standard Python stack (pandas – scipy – statsmodels [16]), data-cleaning and normalization methods (pandas docs [17], Lee [15]), and examples of confidence-interval plotting (Stataiml [18]).

Methodologically, the study combined several complementary stages. First, a theoretical synthesis of classical significance criteria and α-level control based on [1–4], including modern corrections for multiple comparisons (Holm, Bayesian methods). Second, empirical replication of test-statistic distributions via bootstrap and Monte Carlo simulations [2][11], and the evaluation of tail risks alongside Basel-style "traffic-light" VaR back-testing [10][19]. Third, event studies employing the cumulative abnormal return (CAR) methodology in M&A scenarios [9][21], and marketing experiments using A/B tests with a control threshold $p < 0.05$ [8][22]. Fourth, statistical testing of portfolio strategies and diversification at $\alpha = 5\%$ [6][12], accounting for autocorrelation and heteroskedasticity (White's test, Breusch–Pagan). Fifth, analysis of p-value interpretation errors and their consequences based on the survey [5]. Finally, applied implementation of the recommendations was validated by replicating one- and two-sample t-tests, ANOVA, and regression estimates in Python's statsmodels [16], with visualization via fill_between and Matplotlib [18].

RESULTS AND DISCUSSION

In hypothesis testing, the researcher first formulates a null hypothesis, which presumes the "absence of effect," and an alternative hypothesis depicting the anticipated deviation. In finance, this might read: "the strategy's excess return equals zero" versus "the strategy's excess return is positive."

A Type I error represents a false signal: the model is deemed effective despite zero returns. A Type II error is a missed opportunity: a real effect exists, but the test fails to detect it. As Harvey, Liu, and Zhu demonstrated, when hundreds of factors are tested cross-sectionally, the classical threshold $|t| > 2$ ($p \approx 0.05$) generates an avalanche of false discoveries; reliability improves only by requiring $|t| > 3$, corresponding to $p \approx 0.003$, which substantially reduces Type I error risk, albeit at the expense of increasing the chance of overlooking weak but fundamental factors [4].

The p-value merely represents the probability of observing a result at least as extreme, assuming the null hypothesis is true; it neither indicates the likelihood that the hypothesis itself is true nor measures the economic magnitude of the effect. The survey documented in [5], given in Table 1, confirms that this distinction is often ignored. Among 418 Spanish university faculty members, 97% committed at least one typical p-value misinterpretation error, particularly conflating it with the probability of result replication or effect size estimation. Such misconceptions can engender overconfidence in a strategy's "effectiveness" and lead to erroneous investment decisions.

Table 1. Fallacy of the inverse probability [5]

Item	Personality, Evaluation & Psychological Treatments (n = 98)	Behavioral Sciences Methodology (n = 67)	Basic Psychology (n = 56)	Social Psychology (n = 74)	Psychobiology (n = 29)	Developmental & Educational Psychology (n = 94)	Total (n = 418)
The null hypothesis is true	8,2	1,5	7,1	5,4	6,9	12,8	7,4
The null hypothesis is false	65,3	35,8	60,7	66,2	55,2	61,7	58,6
The probability of the null hypothesis has been determined (p = 0,001)	51	58,2	67,9	62,2	62,1	56,4	58,4
The probability of the experimental hypothesis has been deduced (p = 0,001)	40,8	13,4	23,2	36,5	37,9	43,6	33,7
The probability that the null hypothesis is true, given the data obtained, is 0,01	32,7	19,4	25	31,1	41,4	36,2	30,6
% of participants who correctly rate all five statements as false	4,1	19,4	5,4	2,7	0	4,3	6,2

The choice between a one-tailed and two-tailed test depends on whether the theory predicts a specific direction of effect. In the announcement of a buyback, for example, a price increase is anticipated; hence, it is logical to allocate the entire significance level to the upper tail of the distribution, enhancing the sensitivity of a one-tailed test. Conversely, if the sign of the reaction is unknown, such as in initial studies of macroeconomic shocks, a two-tailed criterion is employed, splitting α between both tails. Practical guidelines for event studies in finance emphasize this directional choice based on the hypothesis. Simultaneously, it is essential to remember that the positive autocorrelation of returns—first documented in detail by Lo and MacKinlay for NYSE weekly data—biases standard p-values downward; without adjusting for dependencies, the actual risk of a Type I error remains above the stated level [3].

This bias is particularly evident when assessing active portfolio management. According to Investopedia, fewer than 10% of mutual and closed-end funds exhibit positive alpha relative to the S&P 500 over periods exceeding ten years after fees; consequently, the likelihood of "beating" the benchmark is statistically rare [7]. To decide whether to adopt a strategy claiming 1.7% annual alpha, a manager tests the average quarterly excess return over the risk-free rate; if the sample t-statistic does not exceed 3, the probability that the observed effect is noise outweighs the economic gain from rebalancing, and the strategy is rejected as insufficiently substantiated.

A conservative use of α also extends to marketing experiments: in a project for a major wine distributor, 8–10 A/B tests per month were conducted until 300 conversions per variation were achieved; winning variants produced an average basket increase of +46% and a potential annual revenue uplift of USD 4.2 million, as documented by a $p < 0.05$ report [8].

Cumulative abnormal return models are employed to assess the impact of corporate events. A recent study of Romanian M&A revealed that the aggregate price response in a ten-day window around the announcement exceeded 11% at 1% significance, confirming economic benefits for acquiring shareholders [9].

In market-risk management, α boundaries are enshrined in regulation. Basel's traffic-light framework classifies annual VaR back-test outcomes over 250 days: 0–4 exceptions—green zone; 5–9—yellow; 10 or more—red. The probability of wrongly penalizing a correct model in the green zone does not exceed 11%, while in the red zone, it is below 0.01% [10]. This classification is presented in Table 2.

Table 2. Backtesting Assessment Zones with Exception Counts, Scaling Factor Increases and Cumulative Probabilities [10]

Zone	Number of exceptions	Increase in scaling factor	Cumulative probability
Green Zone	0	0.00	8.11 %
Green Zone	1	0.00	28.58 %
Green Zone	2	0.00	54.32 %
Green Zone	3	0.00	75.81 %
Green Zone	4	0.00	89.22 %
Yellow Zone	5	0.40	95.88 %
Yellow Zone	6	0.50	98.63 %
Yellow Zone	7	0.65	99.60 %
Yellow Zone	8	0.75	99.89 %
Yellow Zone	9	0.85	99.97 %
Red Zone	10 or more	1.00	99.99 %

Predictive-model testing exhibits the same dependence on α. In comparing ARIMA and ARIMA-GARCH for a 120-day forecast of the Chinese index, RMSE declined from 21.73 to 21.66, and all GARCH-component coefficients were significant at the 5% level, justifying the preference for the hybrid model over traditional LSTM approaches, whose errors were substantially larger [11].

From the perspective of strategic diversification, a stringent α aids in determining when the addition of a new asset truly reduces volatility. Empirical data for 41 countries over 1988–2009 indicate that international diversification still delivers significant reductions in return variability, particularly when limited allocations to emerging markets are included; portfolio variance differences are statistically confirmed at the 5% level [12].

In all cases presented, the choice of significance level delineates the boundary between action and inaction. An excessively high α permits costly Type I errors, whereas an overly low α incurs missed opportunities via Type II errors; the optimal balance depends on the economic cost of failure and is justified using the same statistical instruments developed a century ago, now applied to novel data and computational capabilities.

The significance level α establishes a clear demarcation: effects detected above this boundary are interpreted as signals, while those below are treated as noise; primary testing instruments include t-tests and ANOVA. In a one-sample t-test, an investor compares a strategy's average excess return to zero or the risk-free rate: if the sample mean is only 0.84% with volatility of 5.64%, achieving the threshold statistic $t \approx 2$ would require approximately 180 years of monthly observations, rendering the declared "alpha" indistinguishable from random fluctuations and justifying rejection of the strategy [13]. Such calculations enable assessing whether available historical data suffice to sustain the chosen α boundary.

A two-sample t-test is applied to compare two sets of returns. A reverse scenario is observed when contrasting active and index funds: a recent 20-year review [14] reports that 65% of active large-cap products underperformed the S&P 500, and the two-mean comparison t-test formally confirms the absence of persistent active-management outperformance at α = 5%. Across all the above procedures, significance levels are the final barrier between statistical illusion and information suitable for real financial decisions.

Equally important is testing variance constancy. White's universal heteroskedasticity test, based on regression of squared residuals, revealed pronounced heteroskedasticity in the Fama–French five-factor model for Japanese portfolios, invalidating conventional t-values for coefficients and necessitating White-robust standard errors. If the objective is to link variance changes to specific regressors, the Breusch–Pagan test is employed.

Normality testing completes the circle of classical assumptions. For small samples, the Shapiro–Wilk test is most sensitive: a team analyzing daily returns of a diversified portfolio detected a significant departure from normality and heavy tails, rendering VaR estimates overly optimistic. For large data sets, the nonparametric Kolmogorov–Smirnov test is preferable; Lee [15] emphasizes its advantage in assessing empirical versus theoretical return distributions due to tail-form robustness. In both cases, a visual Q-Q plot remains a rapid indicator: noticeable deviations of points from the diagonal preemptively signal that the formal test's p-value is unlikely to support acceptance of the null hypothesis.

Tool selection sets the lower bound on test quality: Python, the most popular data-analysis language, provides the pandas–scipy—stats models stack, where t-tests, F-tests, and bootstrap procedures are executed with a single line; stats models specifically allow formal specification of the linear hypothesis "return = 0" and immediate retrieval of p-values and confidence intervals [16]. Figure 1 illustrates code for testing whether the mean height of university students equals 170 cm using a one-sample t-test.

```python
import pandas as pd
import numpy as np
from scipy import stats

# Generate random heights (mean = 172, std = 5)
data = pd.DataFrame({'heights': np.random.normal(loc=172, scale=5, size=30)})

# Perform one-sample t-test against population mean of 170
t_stat, p_value = stats.ttest_1samp(data['heights'], 170)

print(f"T-statistic: {t_stat}, P-value: {p_value}")
```

Fig. 1. One-Sample t-test code example [16]

Meticulous data preparation is critical, as any error at this stage multiplies the chance of false inference even when α is correctly chosen. The pandas documentation recommends filling missing values in price series via forward/backward fill or date-based interpolation, noting that row deletion introduces bias in seasonal estimates [17]. Practical time-series cleaning guides prescribe a strict workflow: eliminate gaps, detrend, test for stationarity, and only then normalize; following this sequential rule conserves computational resources and reduces the likelihood of spurious autocorrelation. A truncated distribution or winsorization at standard-deviation thresholds is typically applied for outliers in daily returns, preventing economic crises from being misconstrued as statistical anomalies.

Normalization and log transformation are critical when comparing returns of differing periodicities: converting prices to continuous log-returns enhances symmetry, mitigates scale effects, and improves normal approximation, directly increasing the accuracy of confidence intervals at a fixed α. If volatility is deemed time-varying, log-returns are further standardized over a rolling window so that the t-test compares statistics with consistent variance.

Interpretation of results commences with the confidence interval: in Python, it is constructed via stats.t.interval or, for regressions, via summary_frame() in statsmodels; visually, the fill_between function outlines the "plausible" value region, aiding communication with non-statisticians [18]. In multiple-test settings, one reports not individual p-values but their distribution; a violin plot or ECDF curve reveals how many tests fall below the chosen α level, informing the decision to apply Holm or Bayesian corrections before including a factor in a trading system.

As discussed above, the significance level becomes tangible when moving from formulas to real-world decisions. In market-risk management, a bank compares daily P&L against the computed Value-at-Risk. At 95% confidence, regulators expect approximately 12 exceptions over 250 trading days; if exceptions accumulate to ten or more, the model enters Basel's red zone, automatically increasing capital requirements [19].

The same α distinguishes between illusory manager outperformance and genuine alpha. The SPIVA U.S. Scorecard [20] showed that 65% of active large-cap funds underperformed the S&P 500, with a 24-year average failure rate of 64%. At $t \approx 2$, such a frequency could not occur by chance; the statistics confirm that most claimed "selection effects" vanish under a strict 5% threshold.

Event studies offer another example. Following an average M&A announcement, the five-day cumulative abnormal return of the target averaged 1.7%, statistically distinguishable from zero under a two-tailed $\alpha = 5\%$. Thus, merger-arbitrage strategies are supported solely because the p-value surpassed the pre-specified threshold [21].

Business experiments follow the same logic. In testing a discount banner on the Meebox hosting provider's website, conversion rose by 51.9%, average order value by 46.2%, and an A/B test confirmed the entire difference at 98% confidence; had the threshold been more lenient, the effect might have been deemed significant at a more minor increase, but the false-positive risk would have risen sharply [22].

These cases illustrate how a single α value permeates diverse decision levels, from daily VaR monitoring and fund evaluation to A/B campaigns and scoring models. Strict adherence to the threshold reduces costly Type I errors, while tailoring the test to the task—choosing one vs. two-tailed criteria, adjusting for autocorrelation, and multiple comparisons—renders statistical inference a pillar rather than an ornament of the financial plan.

CONCLUSION

This study has demonstrated that the significance level α functions not merely as a statistical parameter but as a pivotal instrument delineating acceptable risk boundaries and the quality of financial decisions. The origins of this concept, rooted in Pearson's χ^2 test and Fisher's practice of $p = 0.05$, established the basis for formal hypothesis verification, while the Neyman–Pearson duality of null and alternative hypotheses introduced control over Type I and Type II errors. The transfer of this methodology to finance via Markowitz's work and the development of market-efficiency

testing afforded a direct monetary interpretation of α, enabling investors to quantitatively relate the probability of a "false" signal to the cost of an erroneous decision.

As the volume and complexity of financial data have grown, traditional asymptotic assumptions have given way to computational procedures—bootstrap and Monte Carlo—necessitating a reevaluation of critical thresholds for heavy-tailed return distributions. Empirical studies of S&P 500 tails and return autocorrelation have affirmed that the classical Gaussian approach underestimates extreme-event probabilities and misjudges real risks. Consequently, adapting α to data structure and multiple testing (Holm and Bayesian corrections, and the requirement of stricter $|t| > 3$) has become essential for reliable asset selection and strategy construction.

Practical examples illustrate the universal application of significance levels: from daily VaR monitoring under Basel's traffic-light rules to the statistical validation of mutual-fund alpha and marketing A/B-test efficacy. In each case, α defines the boundary between signal and noise; selecting one—or two-tailed tests, adjusting for autocorrelation, and correcting for multiple comparisons optimize the trade-off between Type I errors and missed opportunities. Additionally, data preparation and normalization procedures are indispensable: without them, even a correctly chosen α cannot guarantee the validity of inferences.

Thus, applying significance levels in financial planning is a multifaceted process that unites historically established statistical criteria with modern computational methods. Proper α calibration, rigorous data preparation, and attention to empirical market characteristics transform statistical inference from a perfunctory step into a dependable tool for making practical investment decisions.

REFERENCES

1. L. Kennedy-Shaffer, "Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing," *The American Statistician*, vol. 73, no. sup1, pp. 82–90, Mar. 2019, doi: https://doi.org/10.1080/00031305.2018.1537891.
2. P. Gopikrishnan, V. Plerou, L. A. Nunes Amaral, M. Meyer, and H. E. Stanley, "Scaling of the distribution of fluctuations of financial market indices," *Physical Review E*, vol. 60, no. 5, pp. 5305–5316, Nov. 1999, doi: https://doi.org/10.1103/physreve.60.5305.
3. A. W. Lo and A. C. MacKinlay, "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, vol. 1, no. 1, pp. 41–66, Jan. 1988.
4. C. R. Harvey, Y. Liu, and H. Zhu, "… and the Cross-Section of Expected Returns," *Review of Financial Studies*, vol. 29, no. 1, pp. 5–68, Oct. 2016, doi: https://doi.org/10.1093/rfs/hhv059.
5. L. Badenes-Ribera, D. Frías-Navarro, H. Monterde-i-Bort, and M. Pascual-Soler, "Interpretation of the p value: A national survey study in academic psychologists from Spain," *Psicothema*, vol. 27, no. 3, pp. 290–295, 2015, doi: https://doi.org/10.7334/psicothema2014.283.
6. A. Kumar and None Neha, "Testing the Weak Form Efficient Market Hypothesis: An Interpretive Study of Market Efficiency Via Literary Logic and Evidence," *Journal of Global Economics Management and Business Research*, vol. 17, no. 2, pp. 26–36, May 2025, doi: https://doi.org/10.56557/jgembr/2025/v17i29290.
7. J. Chen, "Alpha: Its Meaning in Investing, With Examples," *Investopedia*, Feb. 23, 2024. https://www.investopedia.com/terms/a/alpha.asp (accessed Apr. 15, 2025).
8. Andres Reitsnik, "Case Study: +4.2M Revenue with A/B Testing," *Scandiweb*, Sep. 24, 2024. https://scandiweb.com/blog/case-study-ab-testing-brings-4m-yearly-revenue/ (accessed Apr. 17, 2025).
9. D.-G. Bîlteanu, "Evaluation of the event study in the case of mergers and acquisitions," *Theoretical and Applied Economics*, vol. XXXI, no. 1, pp. 295–312, 2024, Available: https://store.ectap.ro/articole/1737.pdf
10. "Supervisory Framework For The Use Of 'Backtesting' In Conjunction With The Internal Models Approach To Market Risk Capital Requirements," 1996. Available: https://www.bis.org/publ/bcbs22.pdf
11. X. Huang, P. You, X. Gao, and D. Cheng, "Stock Price Prediction Based on ARIMA-GARCH and LSTM," *Atlantis Highlights in Computer Sciences*, pp. 438–448, Jan. 2023, doi: https://doi.org/10.2991/978-94-6463-198-2_45.
12. O. Bouslama and O. B. Ouda, "International Portfolio Diversification Benefits: The Relevance of Emerging Markets," *International Journal of Economics and Finance*, vol. 6, no. 3, Feb. 2014, doi: https://doi.org/10.5539/ijef.v6n3p200.
13. M. Hebner, "Calculations for T-Statistics," *IFA*, Jun. 30, 2024. https://www.ifa.com/articles/calculations_for_t_statistics (accessed Apr. 20, 2025).
14. M. Coleman, "Active Fund Managers vs. Indexes: Analyzing SPIVA Scorecards," *Ifa*, 2019. https://www.ifa.com/articles/spiva-report-active-vs-passive (accessed Apr. 25, 2025).
15. S. Lee, "Kolmogorov-Smirnov Test: 5 Surprising Statistics in Data Analysis," *Number Analytics*, 2025. https://www.numberanalytics.com/blog/kolmogorov-smirnov-test-statistics-analysis (accessed Apr. 28, 2025).
16. A. Yadav, "What is a t-test and when to Use It in Pandas?" *Medium*, Feb. 20, 2025. https://medium.com/%40amit25173/what-is-a-t-test-and-when-to-use-it-in-pandas-9fde303d3c9f (accessed Apr. 27, 2025).

17. "Working with missing data — pandas 0.12.0 documentation," *Pydata*, 2025. https://pandas.pydata.org/pandas-docs/version/0.12.0/missing_data.html (accessed Apr. 26, 2025).

18. "Plotting and Shading Confidence Interval in Python," *Stataiml*, Dec. 29, 2023. https://stataiml.com/posts/plot_ci_python/ (accessed May 01, 2025).

19. "Backtesting VaR: Key Steps and Coverage Tests Explained," *Accounting Insights*, Mar. 05, 2025. https://accountinginsights.org/backtesting-var-key-steps-and-coverage-tests-explained/ (accessed May 03, 2025).

20. A. Ganti, T. Edwards, D. Gioia, F. Chapman, and N. Didio, "SPIVA U.S. Scorecard," S&P Global, 2024. Accessed: May 05, 2025. [Online]. Available: https://www.spglobal.com/spdji/en/documents/spiva/spiva-us-year-end-2024.pdf

21. S. Müller, "Understanding Cumulative Abnormal Return (CAR) in Finance," *Event Study*, Sep. 05, 2024. https://eventstudy.de/blog/cumulative-abnormal-return/ (accessed May 06, 2025).

22. M. Ackerson, "Building the Perfect Landing Page: 37 Conversion Optimization Case Studies," *Growbo*, Nov. 14, 2017. https://www.growbo.com/37-best-conversion-optimization-case-studies/ (accessed May 06, 2025).