



THE COMPENDIUM OF STATISTICAL TESTS

A Comprehensive Collection
of over 250 Statistical Tests

MARUTHY PANNALA, PhD

FIRST EDITION | 2026

The Compendium of Statistical Tests

The Compendium of Statistical Tests

A Comprehensive Collection of Over 250 Statistical Tests

Maruthy Pannala, PhD

Pannala Series

2026

The Compendium of Statistical Tests: A Comprehensive Collection of over 250 Statistical Tests

First Edition

Copyright © 2026 by Maruthy Pannala. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means — electronic, mechanical, photocopying, recording, scanning, or otherwise — without the prior written permission of the publisher, except as permitted under applicable copyright law. Requests for permission should be directed to the publisher at the address below.

Published by

Pannala Series

Greensboro, North Carolina, USA

Email: pannala.series@gmail.com

Website: <https://sites.google.com/view/pannala-series>

Library of Congress Control Number: 2026914047

ISBN: 979-8-9961516-0-8 (Hardcover)

Disclaimer

This book is intended for educational and informational purposes only. While every effort has been made to ensure the accuracy and completeness of the content, the author and publisher make no representations or warranties of any kind, express or implied, with respect to the accuracy, applicability, or completeness of the material herein. The author and publisher shall not be held liable for any errors, omissions, or inaccuracies, nor for any damages or losses — direct, indirect, incidental, or consequential — arising from the use of, or reliance upon, the information contained in this book. Statistical methods should be applied with professional judgment appropriate to the specific context; this book does not constitute professional statistical, legal, financial, or any other form of advice.

Cover design by the author.

Typeset in L^AT_EX using T_EXmaker.

Printed and distributed by IngramSpark.

“What I cannot create, I do not understand.”
— Richard Feynman

Preface

Statistical testing occupies a peculiar position in applied work. On the one hand, statistical tests are foundational tools—embedded deeply in modeling practice, diagnostics, validation, and scientific reporting. On the other hand, they are often invoked mechanically, selected by habit, software defaults, or partial recollection of formal training. The result is not a lack of tests, but a lack of clarity around their correct use, assumptions, interpretation, and limitations.

This book was written to address that gap.

The *Compendium of Statistical Tests* is a single-volume reference work that brings together over 250 statistical tests drawn from across the statistical literature and widely used software ecosystems. These tests are scattered across textbooks, journal articles, technical reports, and software documentation. Practitioners are often required to consult multiple sources—frequently under time pressure—to answer what should be straightforward questions: *Which test is appropriate here? What does it actually measure? What assumptions does it rely on? How should the result be interpreted?*

Each statistical test is presented in a standardized two-page format. The intent is deliberate. A practitioner should be able to open the book at any test and see, at a glance, all information required for correct and informed use. To that end, every test entry follows the same structure, capturing essential elements such as intended use, assumptions, formal test statistic, a plain-English explanation, warnings and best-practice guidance, hypotheses and decision rule, software references, worked example, common pitfalls, and original literature references. The tests are organized into twenty-five chapters, each preceded by a **Conceptual Primer**—a compact, self-contained background section introducing the key ideas, terminology, and mathematical framework needed for the tests that follow.

The selection of tests reflects two guiding principles: **depth** and **breadth**. For commonly encountered diagnostic goals—such as testing for normality, stationarity, or independence—multiple tests are included, each with distinct sensitivities, assumptions, and historical motivations. At the same time, the book spans a wide range of modeling domains: normality testing, general goodness-of-fit testing, classical parametric inference (means, variances, proportions, and rates), nonparametric and rank-based methods, categorical data analysis, correlation and association, regression diagnostics, generalized linear models, mixed models, multivariate analysis, time series analysis (diagnostics, stationarity, cointegration, and volatility), randomness and independence testing, survival and event-time analysis, spatial statistics, quality control and process monitoring, equivalence and non-inferiority testing, robust and resistant methods, permutation and bootstrap inference, outlier detection, and Bayesian inference.

This book is not intended to replace formal statistical training, nor does it attempt to provide full derivations or proofs. Instead, it is designed to serve as a desk-side reference for analysts, model developers, data scientists, econometricians, clinical researchers, quality engineers, environmental scientists, and researchers across disciplines who engage with statistical models and diagnostics as part of their professional work. The emphasis throughout is on correct application, conceptual understanding, and awareness of limitations.

Used as intended, this compendium should help practitioners move beyond rote application of statistical tests and toward more thoughtful, defensible, and transparent analytical decisions.

Acknowledgements

A work of this scope does not emerge in isolation. Many people, over many years, have shaped the thinking behind it.

My parents, Mr. Ramakrishna and Mrs. Indera, gave me the foundation that made everything else possible. The values they instilled—hard work, humility, and intellectual curiosity—have guided me far beyond the pages of this book.

My sister Sailaja has been a steady source of encouragement throughout my life, and this endeavor was no exception. I am grateful for her unwavering support.

My interest in statistics was seeded long before my doctoral work, by the faculty at the Indian Institute of Technology, Kanpur. Professors Debasis Kundu, Divakar Sharma, P. C. Joshi, and G. K. Shukla introduced me to the discipline and laid the statistical foundations I have built upon since. Professor U. B. Tiwari's rigorous instruction in measure theory and real analysis gave me the mathematical discipline underlying all of it. Professor Kundu deserves special mention: fresh from his own doctorate—earned under the supervision of the legendary C. R. Rao—and generous with his time, he encouraged me to pursue a PhD in the United States and helped make that path possible. I owe him more than a sentence can express.

My deepest professional debt belongs to my doctoral supervisor, Dr. K. Krishnamoorthy at the University of Louisiana at Lafayette. Statistics is, at its core, a mathematical discipline—but one largely driven by intuition. Before working with Dr. Krishnamoorthy, I approached the subject almost exclusively through a mathematical lens and, in doing so, missed much of its deeper meaning. His guidance opened a different way of seeing: that ideas come before mathematics, and that equations express intuitions already present. This shift was cemented through the simulation work at the heart of my doctoral dissertation, where I learned to reason about statistical behavior from first principles before writing down a formula. That instinct—to understand what a test is actually doing before asking how it is derived—runs through every page of this book, and is perhaps most visible in the *Plain English* section appearing in every test entry: an attempt to put the idea into words before, and independent of, the mathematics.

My first professional mentor, Dan Walker, helped me understand the business side of quantitative work—that numbers serve decisions, and that communicating them clearly to non-technical audiences is no less important than deriving them correctly. His patience with a young PhD entering the financial industry left a lasting impression.

Over a career spanning more than twenty-five years, I have had the privilege of working alongside many talented and thoughtful colleagues. It is through those interactions—explaining methods, fielding questions, and occasionally bridging the gap between statistical rigor and practical application—that I came to appreciate how much room there is to communicate statistical ideas more clearly across backgrounds and disciplines. That appreciation is, in no small part, what motivated this book.

I am also grateful to my childhood and lifelong friend Sesha Giri, whom I have known since fourth grade. His friendship and steadfast support through every critical juncture in my life have meant more than he probably knows.

Finally, and most importantly, I owe an immeasurable debt to my family. My wife Renukha's support and patience through the long hours this project demanded made it possible to complete. My daughters Soumya and Sneha, my nephew Teja, and my niece Akshara, provided the encouragement and energy that kept me going through the more difficult stretches of this work. My brother-in-law Ravi reviewed the worked examples throughout the book, and my sister-in-law Shobha carefully reviewed the references section—both offering corrections and observations that improved the final result. Their willingness to engage with the material so thoughtfully is something I deeply appreciate.

To all of you—and to Lucky, our Cockapoo, whose company through many late nights was a comfort—thank you.

A Note on Formulas, Variations, and Verification

Statistical knowledge accumulates gradually and sometimes inconsistently. For the majority of well-established tests, software implementations agree closely. But for a meaningful subset—particularly older or more specialized procedures—variations in formulas, finite-sample corrections, and computational approaches exist across both the literature and software. Where such variation exists, the most widely-adopted formulation has been selected. The goal throughout has been accuracy, consistency, and honest acknowledgment of complexity.

Readers should be aware of the following:

- **Approximations have limits.** Many p -value formulas and critical values rely on asymptotic approximations that may behave differently in finite samples. The worked examples and Common Pitfalls sections address these considerations directly.
- **The literature continues to evolve.** While I have endeavored to represent the state of knowledge as of 2026, newer refinements may have since emerged.

Verification. Every formula, worked example, and computational result has been verified through cross-referencing against original source papers and authoritative textbooks, and through independent calculation in Excel, R, Python, and SAS. Despite these efforts, occasional errors may remain. The primary value of this compendium lies in its explanation of test mechanics, assumptions, and proper interpretation—aspects that remain reliable guides regardless of minor formula variations. For critical applications, always verify results using established software and consult original sources when in doubt.

Reporting errors. If you identify any errors or inconsistencies, please contact the author at pannala.series@gmail.com. Confirmed errata will be maintained at <https://sites.google.com/view/pannala-series> and incorporated into future editions.

Statistical literature spans more than a century of scattered, sometimes conflicting contributions. This compendium is one attempt to gather that knowledge into a single accessible reference. It is offered in good faith—and with the hope that, with input from the community, it will improve over time.

— Maruthy Pannala
June 2026

Scope and Classification of Tests

This compendium covers **250+ statistical tests** organized into **25 chapters**, spanning classical inference, regression diagnostics, time series analysis, survival modeling, spatial statistics, and Bayesian procedures. Coverage reflects the diversity of settings in which statistical testing is employed in practice.

Test selection is inherently context-dependent: sample size, data structure, distributional assumptions, robustness requirements, and modeling objectives all influence which procedure is most appropriate. Related tests are therefore grouped together to facilitate comparison across competing approaches. Each chapter opens with a conceptual primer providing the mathematical background needed for the tests that follow; individual entries are presented in a standardized two-page reference format.

The classification scheme is a navigational aid rather than a rigid taxonomy. Many tests admit multiple valid interpretations depending on context, and several appear across more than one chapter. A test's placement should be understood as a primary lens, not a definitive category. Readers are encouraged to use both the table of contents and the Index when locating procedures across related modeling frameworks.

The twenty-five chapters are organized as follows:

1. Normality Tests (NORM)

Tests for consistency with a normal distribution. Includes moment-based tests (skewness, kurtosis), EDF-based procedures, and regression/correlation methods applied to quantile plots, under both simple and composite hypothesis settings.

2. Goodness-of-Fit Tests (GFIT)

Tests evaluating whether data conform to a fully or partially specified distribution. Includes EDF-based statistics (Anderson–Darling, Cramér–von Mises, Kolmogorov–Smirnov, Lilliefors), the chi-square goodness-of-fit test, Kuiper's test & Rao's spacing test for circular data, and Greenwood's test for uniformity of spacings.

3. Means and Location Tests (MEANS)

Parametric inference for comparing population means. Covers one-sample, two-sample, and k -sample settings; pooled and Welch tests; one-way and two-way ANOVA; repeated measures; ANCOVA; and multiple comparison procedures (Tukey, Dunnett, Games–Howell, Scheffé).

4. Variance and Dispersion Tests (VARI)

Tests for equality of variances and dispersion across groups. Ranges from the classical F -test through robust alternatives (Levene, Brown–Forsythe) to nonparametric procedures (Fligner–Killeen, Ansari–Bradley, Mood's, Siegel–Tukey).

5. Tests for Proportions and Rates (PROP)

Inference for binomial proportions and Poisson rates. Covers one-sample proportion tests (z -test, exact binomial, Wilson score), two-sample tests (pooled z -test, Wald, Agresti–Caffo), the Farrington–Manning non-inferiority test, and one- and two-sample Poisson rate tests with exact and asymptotic procedures.

6. Nonparametric and Rank-Based Tests (NPAR)

Distribution-free methods for location and trend inference. Covers two-sample (Mann–Whitney, Mood's median), paired (Wilcoxon signed-rank, sign test), k -sample (Kruskal–Wallis, Friedman), and ordered-alternative tests (Jonckheere–Terpstra, Page's trend test).

7. Correlation and Association Tests (CORR)

Tests for the strength and direction of association across measurement scales. Includes Pearson, Spearman, Kendall's τ , distance correlation, Hoeffding's D , point-biserial, biserial, tetrachoric, polychoric, Cramér's V , partial correlation, and tools for comparing correlations (Fisher's z , Steiger's Z).

8. Categorical Data and Contingency Tables (CTAB)

Tests for count data, proportions, and categorical outcomes. Covers independence and homogeneity (χ^2 , likelihood ratio-G test), exact tests (Fisher's, Barnard's, Boschloo's), paired categorical data (McNemar, McNemar-Bowker, Stuart-Maxwell), repeated binary outcomes (Cochran's Q), stratified analyses (CMH, Breslow-Day, Woolf), trend tests (Cochran-Armitage), and inter-rater agreement (Cohen's κ , Weighted κ , Fleiss' κ , Krippendorff's α , Gwet's AC_1).

9. Linear Regression Diagnostics (REGR)

Specification and diagnostic tests for the classical linear model. Covers heteroscedasticity (Breusch-Pagan, White, Goldfeld-Quandt, Glejser, Harvey-Godfrey), autocorrelation (Durbin-Watson, Breusch-Godfrey), functional form (RESET, Link, Box-Tidwell, Rainbow), structural stability (Chow), multicollinearity (VIF), and observation influence (leverage, Cook's distance, DFFITS, DFBETAS).

10. Generalized Linear Models (GLMS)

Tests for fit, significance, and structural adequacy in GLMs (logistic, Poisson, Gamma, and related families). Covers goodness-of-fit (deviance, Pearson χ^2 , Hosmer-Lemeshow, Le Cessie-van Houwelingen), the LRT/Wald/Score trinity, link function specification (Pregibon), overdispersion (Cameron-Trivedi, Dean's), excess zeros (zero-inflation score test, Vuong), and assumptions specific to multinomial logit (Hausman IIA, Small-Hsiao).

11. Mixed Models (MIXD)

Inference in linear mixed-effects models. Covers likelihood-based tests for random effects (LRT, REML-LRT), score and Wald tests for variance components, the Breusch-Pagan LM test, fixed-effects tests (Wald, Kenward-Roger F), the intraclass correlation coefficient test, and the Hausman specification test.

12. Multivariate Analysis (MULT)

Tests for mean vectors, covariance structure, and distributional form. Covers Hotelling's T^2 (one- and two-sample), MANOVA statistics (Wilks' Λ , Pillai's trace, Hotelling-Lawley trace, Roy's largest root), equality of covariance matrices (Box's M), sphericity (Bartlett, Mauchly, JNS), multivariate normality (Mardia's skewness and kurtosis, Henze-Zirkler), and Benjamini-Hochberg FDR for post-MANOVA multiplicity control.

13. Time Series Diagnostics (TSDG)

Diagnostic tests for time series residuals and forecast evaluation. Covers portmanteau tests for residual autocorrelation (Box-Pierce, Ljung-Box) and the Diebold-Mariano test for comparing predictive accuracy.

14. Stationarity and Unit Root Tests (STAT)

Tests for unit roots and non-stationarity in univariate and panel series. Covers classical tests (DF, ADF, Phillips-Perron), power-enhanced procedures (DF-GLS, Ng-Perron, Schmidt-Phillips), the KPSS stationarity test, structural break tests (Perron, Zivot-Andrews, Clemente-Montañés-Reyes, Lee-Strazicich), seasonal unit root tests (HEGY, Canova-Hansen), the KSS nonlinear unit root test, and panel unit root tests (LLC, IPS) with Pesaran's cross-sectional dependence test.

15. Cointegration Tests (COINT)

Tests for long-run equilibrium among integrated series. Covers residual-based single-equation tests (Engle-Granger, Phillips-Ouliaris), the Johansen system procedures (trace and maximum eigenvalue), break-robust tests (Gregory-Hansen, Maki), alternative single-equation approaches (Shin Residual, Park's added variables), the ARDL bounds test for mixed integration orders, and panel cointegration tests (Pedroni, Westerlund).

16. ARCH and Volatility Tests (ARCH)

Tests for conditional heteroskedasticity in time series residuals. Covers Engle's ARCH LM test, the McLeod-Li test, the Engle-Ng sign and size bias tests for asymmetric volatility, and the ICSS test for structural breaks in variance.

17. Randomness and Independence Tests (RAND)

Tests for patterns, trends, and dependence in ordered sequences. Covers runs-based tests (Wald–Wolfowitz, up-and-down runs), the turning points test, the difference-sign test, monotone trend tests (Mann–Kendall, Cox–Stuart), serial dependence tests (von Neumann ratio, Bartels rank test), Tsay’s test for nonlinear dependence, and the BDS test for general nonlinear dependence.

18. Survival and Event-Time Analysis (SURV)

Tests for time-to-event data with censoring. Covers the log-rank family (Mantel–Cox, Gehan–Breslow, Tarone–Ware, Peto–Peto, Fleming–Harrington), Cox model inference (score test, LRT), proportional hazards diagnostics (Schoenfeld residuals, Grambsch–Therneau), overall adequacy (Cox–Snell residuals), and competing risks methods (Gray’s test, Fine–Gray).

19. Spatial Statistics (SPAT)

Tests for spatial autocorrelation and model specification in areal data. Covers global measures (Moran’s I , Geary’s C , Getis–Ord General G), local statistics (Local Moran’s I /LISA, Getis–Ord G_i^*), the Anselin LM diagnostic sequence (LM-Lag, LM-Error, Robust LM-Lag, Robust LM-Error), and the Join Count test for binary spatial data.

20. Quality Control and Process Monitoring (QUAL)

Statistical process control procedures for stability and capability monitoring. Covers Shewhart chart pattern rules (Western Electric, Nelson), attribute charts (p , np , c , u), memory-based charts (moving average, CUSUM, EWMA), change-point detection (Shiryaev–Roberts, Pettitt’s test), and process capability indices (C_p , C_{pk}).

21. Equivalence and Non-Inferiority Tests (EQVL)

Tests for demonstrating practical equivalence or non-inferiority. Covers the TOST procedure for independent and paired samples (Schuirmann), non-inferiority and superiority-by-margin tests, equivalence testing for proportions, the Mann–Whitney equivalence test, and the Anderson–Hauck test.

22. Robust Statistics (ROBU)

Inferential procedures valid under outliers, heavy tails, and non-normality. Includes trimmed mean tests (Yuen’s one-sample, two-sample, and ANOVA extensions), the Winsorised variance test, Welch–James test, Brunner–Munzel test, percentile bootstrap inference, Huber’s M -estimator test, and the Theil–Sen slope estimator and test.

23. Permutation and Randomization Tests (PERM)

Exact and approximate inference from resampling. Covers permutation t -test and ANOVA, permutation test for correlation, Mantel matrix permutation test, randomization tests, bootstrap hypothesis tests, and bootstrap confidence interval tests.

24. Outlier Detection Tests (OUTL)

Formal tests for anomalous observations. Covers single-outlier parametric tests (Grubbs, Dixon’s Q , Chauvenet’s criterion), multiple-outlier procedures (Rosner’s generalized ESD, Tietjen–Moore, Peirce’s criterion), robust univariate detection (Iglewicz–Hoaglin modified z -score), and multivariate detection via Mahalanobis distance (classical and MCD-based).

25. Bayesian Analysis (BAYES)

Bayesian hypothesis testing and model comparison. Covers the Bayes Factor (BF_{10}) and the Savage–Dickey density ratio, Bayesian model comparison and averaging, Bayesian analogues of classical tests (t -test, ANOVA, correlation, regression, proportion), the HDI + ROPE framework for practical significance, and the posterior predictive p -value for model checking.

How to Use This Compendium

This book is designed as a working reference rather than a sequential read. Navigate directly to the relevant chapter based on your analytical task. Within each chapter, multiple tests addressing the same goal are presented side by side, making comparison straightforward.

Each chapter opens with a **Conceptual Primer**—a compact introduction to the key ideas, terminology, and mathematical framework needed for the tests that follow. Readers already familiar with the subject can proceed directly to the individual entries.

Anatomy of a Test Entry

Each test occupies a standardized two-page format. The elements are as follows:

Element	Content
Title & Subtitle	Name and primary purpose of the test.
Metadata Line	Classification code (e.g. NORM.1, REGR.7), Category, data type, null distribution.
When to Use	Scope, required data type, and inferential question answered. Guides selection and prevents misuse.
Assumptions	Explicitly stated conditions. Violations may invalidate the test or alter its null distribution.
Test Statistic	Formal mathematical definition in standard notation.
Plain English	Notation-free explanation of what the test measures and how departures from the null manifest in practice.
Warning	Practical cautions, finite-sample sensitivities, and modern alternatives where relevant.
Hypotheses & Decision Rule	Null and alternative hypotheses in both mathematical and/or verbal form, with the decision rule.
Software	Representative code in Python , R , and SAS . Code shown is illustrative; extended implementations are planned for https://sites.google.com/view/pannala-series .
Additional Notes	Supplementary technical material: derivations, asymptotic theory, connections to related tests. Skippable without loss of continuity.
Miscellanea	Historical background, etymology, and anecdotes.
References	Up to four or five key citations as a starting point for deeper study.
Worked Example	Complete numerical illustration from raw data to conclusion.
Common Pitfalls	Frequent errors in application or interpretation: assumption misuse, confusion with related tests, misreading of non-rejection.
Additional Material	Supplementary formulas or technical notes, included where warranted.

The uniform layout is intentional. After a few entries, readers will internalize the structure and extract what they need rapidly—whether the formula, the decision rule, the worked example, or the plain-English explanation. The *When to Use*, *Assumptions*, and *Warning* sections together constitute a decision-support framework embedded within every entry.

Contents

Preface	iii
Acknowledgements	v
A Note on Formulas, Variations, and Verification	vii
Scope and Classification of Tests	ix
How to Use This Compendium	xiii
Notations and Abbreviations	xxv
Part 0: Inference Foundations	1
Introduction	2
Part I: Normality	7
A Conceptual Primer	8
[NORM.1]: Shapiro-Wilk Test.	10
[NORM.2]: Shapiro-Francia Test	12
[NORM.3]: Anderson–Darling Test.	14
[NORM.4]: Cramér–von Mises Test	16
[NORM.5]: Kolmogorov–Smirnov Test (Lilliefors Test)	18
[NORM.6]: Jarque-Bera Test	20
[NORM.7]: D’Agostino–Pearson K^2 Test	22
Raw Data & Charts	24
Technical Notes.	26
Part II: General Goodness-of-Fit	27
A Conceptual Primer	28
[GFIT.1]: Anderson–Darling Test.	30
[GFIT.2]: Cramér–von Mises Test	32
[GFIT.3]: Kolmogorov–Smirnov Test (Two-Sample).	34
[GFIT.4]: Lilliefors Test for Exponential Distribution	36
[GFIT.5]: Kuiper’s Test	38
[GFIT.6]: Rao’s Spacing Test	40
[GFIT.7]: Greenwood Test for Uniformity	42
[GFIT.8]: Chi-Square Goodness-of-Fit Test	44
Part III: Means and Locations	47
A Conceptual Primer	48
[MEAN.1]: Z-Test (One-Sample)	50
[MEAN.2]: One-Sample t-Test	52
[MEAN.3]: Z-Test (Two-Sample)	54
[MEAN.4]: Two-Sample t-Test (Independent).	56
[MEAN.5]: Welch’s t-Test	58
[MEAN.6]: Paired t-Test	60
[MEAN.7]: One-Way ANOVA F Test	62
[MEAN.8]: Welch’s ANOVA	64

[MEAN.9]: Two-Way ANOVA	66
[MEAN.10]: Repeated Measures ANOVA	68
[MEAN.11]: ANCOVA F Test	70
[MEAN.12]: Tukey's HSD Test	72
[MEAN.13]: Dunnett's Test	74
[MEAN.14]: Games-Howell Test	76
[MEAN.15]: Scheffé's Test	78
Technical Notes	80
Part IV: Variances and Scale	85
A Conceptual Primer	86
[VARI.1]: F-Test for Equality of Two Variances	88
[VARI.2]: Bartlett's Test	90
[VARI.3]: Hartley's F_{max} Test	92
[VARI.4]: Levene's Test	94
[VARI.5]: Brown-Forsythe Test	96
[VARI.6]: Mood's Dispersion Test	98
[VARI.7]: Ansari-Bradley Test	100
[VARI.8]: Siegel-Tukey Test	102
[VARI.9]: Fligner-Killeen Test	104
Part V: Proportions and Rates	107
A Conceptual Primer	108
[PROP.1]: One-Sample Proportion Test (Wald)	110
[PROP.2]: Exact Binomial Test	112
[PROP.3]: Score Test (Wilson) for Proportions	114
[PROP.4]: Two-Sample Proportion Test (Pooled Wald Z-test)	116
[PROP.5]: Wald Test for Proportions (General k -Sample)	118
[PROP.6]: Farrington-Manning Test	120
[PROP.7]: One-Sample Poisson Rate Test	122
[PROP.8]: Two-Sample Poisson Rate Test	124
Part VI: Nonparametric Statistics	127
A Conceptual Primer	128
[NPAR.1]: Sign Test	130
[NPAR.2]: Wilcoxon Signed-Rank Test	132
[NPAR.3]: Mann-Whitney U Test	134
[NPAR.4]: Mood's Median Test	136
[NPAR.5]: Kruskal-Wallis Test	138
[NPAR.6]: Friedman Test	140
[NPAR.7]: Jonckheere-Terpstra Test	142
[NPAR.8]: Page's Trend Test	144
Special Topic: Quade's Test	146
Corrections for Ties	148

Part VII: Correlation and Association	149
A Conceptual Primer	150
[CORR.1]: Pearson's Correlation Test	152
[CORR.2]: Spearman's Rank Correlation Test	154
[CORR.3]: Kendall's τ Test	156
[CORR.4]: Distance Correlation Test.	158
[CORR.5]: Hoeffding's D Test	160
[CORR.6]: Point-Biserial Correlation Test	162
[CORR.7]: Biserial Correlation	164
[CORR.8]: Tetrachoric Correlation Test	166
[CORR.9]: Polychoric Correlation	168
[CORR.10]: Cramér's V	170
[CORR.11]: Partial Correlation Test	172
[CORR.12]: Fisher's Z Transformation Test.	174
[CORR.13]: Steiger's Z Test	176
Technical Notes.	178
Part VIII: Categorical Data and Contingency Tables	179
A Conceptual Primer	180
[CTAB.1]: Chi-Square Test of Independence	186
[CTAB.2]: G-Test of Independence	188
[CTAB.3]: Fisher's Exact Test.	190
[CTAB.4]: Barnard's Exact Test.	192
[CTAB.5]: Boschloo's Test.	194
[CTAB.6]: McNemar's Test	196
[CTAB.7]: McNemar-Bowker Test	198
[CTAB.8]: Stuart-Maxwell Test	200
[CTAB.9]: Cochran's Q Test	202
[CTAB.10]: Cochran-Mantel-Haenszel Test	204
[CTAB.11]: Breslow-Day Test.	206
[CTAB.12]: Woolf's Test	208
[CTAB.13]: Cochran-Armitage Test	210
[CTAB.14]: Cohen's Kappa	212
[CTAB.15]: Weighted Kappa	214
[CTAB.16]: Fleiss' Kappa	216
[CTAB.17]: Krippendorff's Alpha	218
[CTAB.18]: Gwet's AC_1	220
Raw Data	222
Worked Example: Mantel-Haenszel's Pooled Estimator.	223
Worked Example: CMH's Row Mean Scores Variant	224
Worked Example: CMH's Correlation Variant	225

Part IX: Regression Diagnostics	227
A Conceptual Primer	228
[REGR.1]: Breusch-Pagan Test	232
[REGR.2]: White's Test.	234
[REGR.3]: Goldfeld-Quandt Test	236
[REGR.4]: Glejser Test	238
[REGR.5]: Harvey-Godfrey Test	240
[REGR.6]: Durbin-Watson Test	242
[REGR.7]: Breusch-Godfrey Test	244
[REGR.8]: Ramsey's RESET Test	246
[REGR.9]: Link Test	248
[REGR.10]: Box-Tidwell Test	250
[REGR.11]: Rainbow Test	252
[REGR.12]: Chow Test	254
[REGR.13]: VIF Test	256
[REGR.14]: Leverage (Hat Values) Test	258
[REGR.15]: Cook's Distance Test.	260
[REGR.16]: DFFITS Test	262
[REGR.17]: DFBETAS.	264
Technical Notes.	266
Part X: Generalized Linear Models	267
A Conceptual Primer	268
[GLMS.1]: Deviance Goodness-of-Fit Test	272
[GLMS.2]: Pearson Chi-Square Goodness-of-Fit Test	274
[GLMS.3]: Hosmer-Lemeshow Test	276
[GLMS.4]: Le Cessie-van Houwelingen Test	278
[GLMS.5]: Likelihood Ratio Test (LRT).	280
[GLMS.6]: Wald Test	282
[GLMS.7]: Score Test (Rao's)	284
[GLMS.8]: Link Function Test (Pregibon).	286
[GLMS.9]: Overdispersion Test (Cameron-Trivedi)	288
[GLMS.10]: Dean's Test	290
[GLMS.11]: Zero-Inflation Score Test	292
[GLMS.12]: Vuong Test	294
[GLMS.13]: Hausman Test for IIA	296
[GLMS.14]: Small-Hsiao Test	298
Raw Data	300
Technical Notes.	301
Likelihoods, MLEs, and Key Quantities	303

Part XI: Mixed/Hierarchical Models	305
A Conceptual Primer	306
[MIXD.1]: Likelihood Ratio Test for Random Effects	308
[MIXD.2]: Restricted Likelihood Ratio Test (REML)	310
[MIXD.3]: Breusch-Pagan LM Test (Random Effects)	312
[MIXD.4]: Score Test for Variance Components	314
[MIXD.5]: Wald Test for Variance Components	316
[MIXD.6]: Wald Test for Fixed Effects	318
[MIXD.7]: F-Test with Kenward-Roger Approximation	320
[MIXD.8]: Intraclass Correlation Coefficient Test	322
[MIXD.9]: Hausman Specification Test	324
Part XII: Multivariate Analysis	327
A Conceptual Primer	328
[MULT.1]: Hotelling's T^2 Test (One-Sample).	332
[MULT.2]: Hotelling's T^2 Test (Two-Sample)	334
[MULT.3]: Wilks' Lambda Test	336
[MULT.4]: Pillai's Trace Test	338
[MULT.5]: Hotelling-Lawley Trace Test	340
[MULT.6]: Roy's Largest Root Test	342
[MULT.7]: Box's M Test	344
[MULT.8]: Bartlett's Test of Sphericity	346
[MULT.9]: Mauchly's Test of Sphericity.	348
[MULT.10]: John-Nagao-Sugiura Test for Sphericity	350
[MULT.11]: Mardia's Test (Multivariate Normality).	352
[MULT.12]: Henze-Zirkler Test	354
[MULT.13]: Benjamini-Hochberg (FDR)	356
Technical Notes.	358
Part XIII: Time Series Diagnostics	361
A Conceptual Primer	362
[TSDG.1]: Portmanteau Test.	364
[TSDG.2]: Ljung-Box Test	366
[TSDG.3]: Diebold-Mariano Test	368
Raw Data & Charts	370
Part XIV: Stationarity and Unit Roots	371
A Conceptual Primer	372
[STAT.1]: Dickey-Fuller (DF) Test	374
[STAT.2]: Augmented Dickey-Fuller (ADF) Test.	376
[STAT.3]: Phillips-Perron (PP) Test.	378
[STAT.4]: DF-GLS Test.	380
[STAT.5]: Ng-Perron Test	382
[STAT.6]: Schmidt-Phillips Test.	384

[STAT.7]: KPSS Test	386
[STAT.8]: Perron Test	388
[STAT.9]: Zivot-Andrews Test	390
[STAT.10]: Clemente-Montañés-Reyes Test	392
[STAT.11]: Lee-Strazicich Test	394
[STAT.12]: HEGY Test	396
[STAT.13]: Canova-Hansen Test	398
[STAT.14]: Kapetanios-Shin-Snell Test	400
[STAT.15]: Levin-Lin-Chu Test	402
[STAT.16]: Im-Pesaran-Shin (IPS) Test	404
[STAT.17]: Pesaran's CD Test	406
Raw Data	408
Technical Notes.	409
Part XV: Cointegration	411
A Conceptual Primer	412
[COINT.1]: Engle-Granger Two-Step Test	416
[COINT.2]: Phillips-Ouliaris Test	418
[COINT.3]: Johansen Trace Test	420
[COINT.4]: Johansen Maximum Eigenvalue Test	422
[COINT.5]: Gregory-Hansen Test (Single Break)	424
[COINT.6]: Maki Test (Multiple Breaks)	426
[COINT.7]: Shin Residual Cointegration Test.	428
[COINT.8]: Park Added Variables Test.	430
[COINT.9]: ARDL Bounds Test.	432
[COINT.10]: Pedroni Test (Panel).	434
[COINT.11]: Westerlund Test (Panel)	436
Raw Data	438
Part XVI: Volatility and Conditional Heteroskedasticity	439
A Conceptual Primer	440
[ARCH.1]: ARCH Test (Engle's LM Test).	442
[ARCH.2]: McLeod-Li Test	444
[ARCH.3]: Engle-Ng Sign Bias Test	446
[ARCH.4]: ICSS Test	448
Raw Data	450
Part XVII: Randomness and Independence	451
A Conceptual Primer	452
[RAND.1]: Wald-Wolfowitz Runs Test.	456
[RAND.2]: Up-and-Down Runs Test	458
[RAND.3]: Turning Points Test	460
[RAND.4]: Difference-Sign Test	462
[RAND.5]: Mann-Kendall Test	464

[RAND.6]: Cox-Stuart Test	466
[RAND.7]: Von Neumann's Ratio Test.	468
[RAND.8]: Bartels Test.	470
[RAND.9]: Tsay's Test	472
[RAND.10]: BDS Test	474
Detailed Worked Example: BDS Test	476
Raw Data	479
Technical Notes.	480
Part XVIII: Survival Analysis	485
A Conceptual Primer	486
[SURV.1]: Log-Rank Test (Mantel-Cox).	490
[SURV.2]: Gehan-Breslow Test (Wilcoxon).	492
[SURV.3]: Tarone-Ware Test	494
[SURV.4]: Peto-Peto Test.	496
[SURV.5]: Fleming-Harrington Test	498
[SURV.6]: Cox Proportional Hazards Score Test	500
[SURV.7]: Likelihood Ratio Test (Cox Model)	502
[SURV.8]: Schoenfeld Residuals Test.	504
[SURV.9]: Grambsch-Therneau Test	506
[SURV.10]: Cox-Snell Residuals Test	508
[SURV.11]: Gray's Test (Competing Risks)	510
[SURV.12]: Fine-Gray Test.	512
Raw Data	514
Special Topic: MaxCombo Test	515
Survival Analysis Demystified	518
Part XIX: Spatial Statistics	519
A Conceptual Primer	520
[SPAT.1]: Moran's I Test	522
[SPAT.2]: Geary's C Test	524
[SPAT.3]: Getis-Ord General G Test	526
[SPAT.4]: Local Moran's I (LISA)	528
[SPAT.5]: Getis-Ord G_i^* Test.	530
[SPAT.6]: LM Test for Spatial Lag	532
[SPAT.7]: LM Test for Spatial Error	534
[SPAT.8]: Robust LM Test (Spatial Lag).	536
[SPAT.9]: Robust LM Test (Spatial Error).	538
[SPAT.10]: Join Count Test.	540
Part XX: Quality Control and Process Monitoring	543
A Conceptual Primer	544
[QUAL.1]: Pattern Recognition Rules (Western Electric & Nelson)	546
[QUAL.2]: Attribute Control Chart Tests	548

[QUAL.3]: Moving Average Control Chart Test	550
[QUAL.4]: CUSUM Test	552
[QUAL.5]: EWMA Test	554
[QUAL.6]: Shiryaev-Roberts Test	556
[QUAL.7]: Change-Point Detection Test (Pettitt's)	558
[QUAL.8]: Process Capability Test (Cp, Cpk)	560
Part XXI: Equivalence and Non-Inferiority	563
A Conceptual Primer	564
[EQVL.1]: TOST for Independent Samples (Schuirmann's Test)	566
[EQVL.2]: TOST (Paired Data)	568
[EQVL.3]: Non-Inferiority Test	570
[EQVL.4]: Superiority-by-Margin Test	572
[EQVL.5]: Equivalence Test for Proportions	574
[EQVL.6]: Mann-Whitney Equivalence Test	576
[EQVL.7]: Anderson-Hauck Test	578
Part XXII: Robust and Resistant Statistics	581
A Conceptual Primer	582
[ROBU.1]: Yuen's Test (One-Sample)	586
[ROBU.2]: Yuen's Test (Two-Sample)	588
[ROBU.3]: Trimmed Mean ANOVA	590
[ROBU.4]: Winsorized Variance Test	592
[ROBU.5]: Welch-James Test	594
[ROBU.6]: Brunner-Munzel Test	596
[ROBU.7]: Percentile Bootstrap Test	598
[ROBU.8]: Huber's M-Test	600
[ROBU.9]: Theil-Sen Slope Test	602
Special Topic: ANOVA-Type Statistic (ATS)	604
Special Topic: Aligned Rank Transform (ART)	606
Part XXIII: Permutation and Resampling	609
A Conceptual Primer	610
[PERM.1]: Permutation t-Test	612
[PERM.2]: Permutation ANOVA	614
[PERM.3]: Permutation Test for Correlation	616
[PERM.4]: Mantel Test	618
[PERM.5]: Randomization Test	620
[PERM.6]: Bootstrap Hypothesis Test	622
[PERM.7]: Bootstrap Confidence Interval Test	624
Detailed Worked Example: Bootstrap Confidence Intervals	626
Technical Notes	630

Part XXIV: Outlier Detection	631
A Conceptual Primer	632
[OUTL.1]: Grubbs' Test	634
[OUTL.2]: Dixon's Q Test	636
[OUTL.3]: Chauvenet's Criterion	638
[OUTL.4]: Rosner's Test	640
[OUTL.5]: Tietjen-Moore Test	642
[OUTL.6]: Peirce's Criterion	644
[OUTL.7]: Iglewicz and Hoaglin's Test	646
[OUTL.8]: Mahalanobis Distance Test	648
Technical Notes	650
Part XXV: Bayesian Analysis	653
A Conceptual Primer	654
[BAYES.1]: Bayes Factor Test (BF_{10})	658
[BAYES.2]: Savage-Dickey Density Ratio Test	660
[BAYES.3]: Bayesian Model Comparison	662
[BAYES.4]: Bayesian t-Test	664
[BAYES.5]: Bayesian ANOVA	666
[BAYES.6]: Bayesian Correlation Test	668
[BAYES.7]: Bayesian Linear Regression Test	670
[BAYES.8]: Bayesian Proportion Test	672
[BAYES.9]: HDI + ROPE Test (Combined Approach)	674
[BAYES.10]: Posterior Predictive p-Value Test	676
Bayesian Workflow: An Instructive Example	678
Companion Guide to Worked Examples	681
Appendix	685
Probability Theory: A Practical Reference	687
Statistical Foundations: A Practical Reference	691
Statistical Distributions: Quick Reference	695
References	705
Index	725
About the Author	739
Errata and Corrections	740

Notations and Abbreviations

The following notations and abbreviations are used consistently throughout this compendium. Where a symbol serves double duty (e.g., p as both p-value and dimension), the meaning is clear from context and noted below. $\log$ denotes the natural logarithm (base e) unless explicitly stated otherwise.

Greek Letters and Mathematical Symbols

α	Significance level (Type I error)
β	Type II error; regression coefficient
ϵ, η	Sphericity correction (Greenhouse–Geisser, Huynh–Feldt)
λ	Likelihood ratio statistic; Poisson rate
μ	Population mean
σ^2	Population variance
σ	Population standard deviation
τ	Kendall’s rank correlation
χ^2_ν	Chi-square, ν degrees of freedom
$\Phi(\cdot)$	Standard normal CDF
$\phi(\cdot)$	Standard normal PDF
Σ	Population covariance matrix
θ	Generic parameter (or vector)
$\hat{\theta}$	MLE or point estimate of θ

Roman Symbols and Matrix Notation

F_{ν_1, ν_2}	F -distribution, ν_1 and ν_2 df
H_0	Null hypothesis
H_1	Alternative hypothesis
$\mathbf{I}_p$	$p \times p$ identity matrix
$\mathbf{S}$	Sample covariance matrix
$L(\theta)$	Likelihood function at θ
$\ell(\theta)$	Log-likelihood: $\log L(\theta)$
n	Sample size
$N(\mu, \sigma^2)$	Normal(μ, σ^2) distribution
p	p-value <i>or</i> dimension (context)
$\mathbb{R}^p$	p -dimensional Euclidean space
t_ν	t -distribution, ν df
$\text{tr}(\cdot)$	Matrix trace
$z_\alpha, z_{1-\alpha}$	Standard normal critical values (tail direction is context-dependent)

Abbreviations

AIC	Akaike Information Criterion		
ANOVA	Analysis of Variance	LOO	Leave-One-Out
ANCOVA	Analysis of Covariance	LR	Likelihood Ratio
ARL	Average Run Length	LRT	Likelihood Ratio Test
ARL ₀	In-control Average Run Length	MANOVA	Multivariate Analysis of Variance
ARL ₁	Out-of-control Average Run Length	MIC	Maximal Information Coefficient
BIC (SIC)	Bayesian (Schwarz) Information Criterion	MLE	Maximum Likelihood Estimator
CDF	Cumulative Distribution Function	MR	Mean Reversion
CV	Critical Value	MSE	Mean Squared Error
df	Degrees of freedom	MVN	Multivariate Normal distribution
DGP	Data Generating Process	NB	Negative Binomial distribution
FDR	False Discovery Rate	OLS	Ordinary Least Squares
FTR	Failed to Reject H_0	PDF	Probability Density Function
FWER	Family-Wise Error Rate	PH	Proportional Hazards
GLM	Generalized Linear Model	PMF	Probability Mass Function
GLMM	Generalized Linear Mixed Model	PSIS	Pareto Smoothed Importance Sampling
GLS	Generalized Least Squares	QC	Quality Control
GOF	Goodness of Fit	SD	Standard Deviation
HAC	Heteroscedasticity & Autocorrelation Consistent	SE	Standard Error
i.i.d. (IID)	Independent and Identically Distributed	SSCP	Sums of Squares and Cross-products
IIA	Independence of Irrelevant Alternatives	2SLS	Two-Stage Least Squares
IV	Instrumental Variables	ZIP	Zero-Inflated Poisson model
LM	Lagrange Multiplier (score test)	ZINB	Zero-Inflated Negative Binomial model

Distributions: N , χ^2 , F , t denote Normal, Chi-square, F -, and Student's t -distributions.

Software: R packages as `pkg::fn()`; SAS procedures as `PROC NAME`.

PART 0

Foundations of Statistical Inference

Core principles for interpreting evidence, uncertainty, and decisions

INFERENCE FOUNDATIONS

INTRO - Foundations of Statistical Inference

Before delving into the tests themselves, this chapter builds the conceptual scaffolding that unifies every test in this compendium. Understanding these foundations transforms a collection of formulas into a coherent system for reasoning under uncertainty.

I. The Architecture of Inference: Signal, Noise, and Extraction

Statistical testing rests on a simple but profound idea: every empirical measurement blends **signal**—the underlying structure we seek—with **noise**, the random variation we must overcome to detect it.

1. All data are mixtures. Any observable Y can be expressed as

$$Y = f(X) + \varepsilon,$$

where $f(X)$ is the systematic component and ε the stochastic residue. Inference is difficult precisely when the signal-to-noise ratio is low: large noise can obscure meaningful differences, while high signal can make even small samples persuasive.

2. Models as filters. Statistical models act as filters separating structure from randomness. Estimation methods (OLS, MLE) extract the systematic component; standard errors quantify uncertainty due to noise. Hypothesis tests compare signal to noise through ratios such as

$$t = \frac{\text{signal}}{\text{noise}}.$$

A result becomes “statistically significant” only when the estimated signal rises clearly above background noise.

3. Design as a noise-reduction tool. Good experimental design improves inference not by changing the truth but by reducing uncertainty. Paired designs subtract subject-level variability; ANCOVA removes covariate effects. In both cases, the smaller noise in the denominator sharpens detection of the signal in the numerator.

4. When noise masquerades as signal. Overfitting occurs when a model becomes too flexible and begins to chase accidental patterns. Such models fit the current sample perfectly but generalize poorly. Regularization, cross-validation, parsimony, and diagnostic checking ensure that we learn the signal, not the static (noise).

Key insight: Every test in this compendium is a specific method for comparing signal to noise. The art of inference lies not merely in collecting more data, but in *shaping* data and models so that meaningful patterns emerge cleanly from background variability.

II. The Inferential Chain

Every test typically follows the same five-step logic:

- Model:** formulate a probability model for the data-generating process.
- Null hypothesis H_0 :** specify a restricted version of the model (no effect, no difference, no association).
- Test statistic:** compute a scalar summary that measures deviation from H_0 .
- p -value:** the probability, under H_0 , of a statistic at least as extreme as the one observed.
- Decision:** reject H_0 if $p \leq \alpha$; otherwise fail to reject it.

The decision is probabilistic—not a deterministic verdict about truth. Statistical significance reflects incompatibility with a model, not practical importance.

Direction of rejection: The rejection region (sometimes referred to as critical region) follows the direction of the alternative hypothesis. For right-tailed tests ($H_1 : \theta > \theta_0$), reject for large positive values ($Z > z_\alpha$); for left-tailed tests ($H_1 : \theta < \theta_0$), reject for large negative values ($Z < -z_\alpha$); and for two-sided tests ($H_1 : \theta \neq \theta_0$), reject for large deviations in either direction ($|Z| > z_{\alpha/2}$). In non-inferiority testing, the null asserts that the new treatment falls short by more than an acceptable margin δ : $H_0 : \theta \leq \theta_0 - \delta$; reject if $Z > z_\alpha$, where Z is computed against the margin boundary $\theta_0 - \delta$, not the reference value θ_0 itself. For equivalence, two one-sided tests (TOST) are used: reject $H_{01} : \theta \leq \theta_L$ using $Z_L = (\hat{\theta} - \theta_L)/SE > z_\alpha$, and separately reject $H_{02} : \theta \geq \theta_U$ using $Z_U = (\hat{\theta} - \theta_U)/SE < -z_\alpha$; conclude equivalence only if both nulls are rejected simultaneously.

III. Type I and Type II Errors

Every test balances two types of errors.

	H_0 True	H_0 False
Reject H_0	FP: Type I (α)	TP: Power ($1-\beta$)
Fail to reject H_0	TN: Correct ($1-\alpha$)	FN: Type II (β)

Here, TP-True Positive, FP-False Positive, TN-True Negative, and FN-False Negative.

- α : probability of a false positive (rejecting when H_0 is true).
- β : probability of a false negative (missing a true effect).
- Power = $1 - \beta$: probability of correctly detecting a true effect.

Conservative vs. liberal tests: A test is *conservative* if the actual Type I error $< \alpha$ (fewer false positives, lower power); *liberal* if the actual Type I error $> \alpha$ (inflated false positives).

Key metrics from the confusion matrix:

$$\text{Recall} = \frac{TP}{TP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}.$$

Precision measures *how accurate positive predictions are*; **Recall** measures *how many positives are captured*.

The fundamental trade-off: for fixed n , lowering α (Type I) raises β (Type II). The only way to reduce both simultaneously is to increase the sample size or the effect size. Effect size (δ) quantifies the magnitude of departure from H_0 under H_1 . Power depends on δ , n , α , and test directionality.

Warning: “Fail to reject” means *insufficient evidence*—not that H_0 is true. A study may simply be underpowered.

IV. Why Classical Tests Never Prove H_0

In classical hypothesis testing, we either *reject* H_0 or *fail to reject* it—never *accept* it. This is not pedantic language; it reflects a fundamental asymmetry in how statistical evidence works.

The direction of evidence. Tests are calibrated to answer one question:

If H_0 were true, how surprising would the observed data be?

A small p -value indicates that the data are highly inconsistent with H_0 , providing grounds to reject it. A large p -value indicates that the data are *not sufficiently incompatible* with H_0 —it does not confirm H_0 .

Absence of evidence is not evidence of absence (*Quote attributed to Carl Sagan*). Failing to reject H_0 may reflect limited sample size, low power, high variability, or an insensitive test. The data may simply be uninformative.

The legal analogy. A defendant is found *guilty* or *not guilty*—never “innocent.” A not-guilty verdict reflects insufficient evidence for conviction, not proof of innocence. Likewise, failing to reject H_0 reflects insufficient evidence *against* it, not proof *of* it.

What this means in practice:

- Non-rejection does not validate modeling assumptions.
- Large p -values do not imply equivalence.
- Power, effect size, and diagnostic plots remain essential regardless of the p -value.

Warning: “No evidence against” $H_0 \neq$ “evidence for” H_0 . For methods that *can* establish equivalence, see the EQVL chapter.

V. The p -Value: Definition and Misuse

Definition:

$$p = P(T \geq t_{\text{obs}} \mid H_0).$$

What the p -value is:

- A measure of *compatibility* between the data and H_0 .
- Sensitive to both effect size and n —a large n can make trivial effects produce small p -values.
- A continuous measure of evidence, not a binary gate: $p = 0.049$ and $p = 0.051$ carry essentially the same information.

What the p -value is not:

- $P(H_0 \mid \text{data})$ —the probability H_0 is true (Bayesian concept; requires a prior; see § VII).
- The probability that results are “due to chance.”
- A measure of effect size or practical importance.

ASA Statement (2016): $p < 0.05$ alone does not constitute scientific evidence. Conclusions require effect sizes, confidence intervals, replication, and subject-matter reasoning—not just a threshold crossing. (ASA: American Statistical Association; Wasserstein & Lazar (2016))

VI. The Symmetry of Evidence: p -Value Uniformity

Beneath the diversity of tests in this compendium lies a fundamental mathematical symmetry governed by the **Probability Integral Transform (PIT)**.

Under a correctly specified H_0 , any *continuous* test statistic T maps via its CDF F_T to a uniform distribution:

$$p = 1 - F_T(T) \sim U(0, 1) \quad \text{under } H_0.$$

This means $P(p \leq \alpha \mid H_0) = \alpha$ for every $\alpha \in [0, 1]$ —the calibration principle underlying every classical test.

Uniformity as a diagnostic tool. When running a battery of tests (e.g., in meta-analysis), the distribution of p -values is itself informative (e.g: see GFIT.7):

- **Clustering near 0:** genuine effects are present; the data contain signal beyond what is expected under the null.
- **Clustering near 1:** conservative procedures, over-correction for multiplicity, or model misspecification producing systematically large p -values.
- **Non-uniform pattern:** calibration failure, violated assumptions, publication bias, or data snooping.

A single significant p -value may be a fluke; the *distribution* of p -values across independent tests provides a signature of statistical integrity. A histogram that departs substantially from flatness is often an early warning sign that assumptions have been violated or that systematic effects remain unmodeled.

This observation underlies many modern procedures, including false discovery rate (FDR) control and empirical Bayes methods, which model the distribution of p -values directly.

Technical caveats: exact uniformity requires (i) continuous test statistics, (ii) simple null hypotheses, and (iii) independent tests. Discrete tests (Fisher’s exact, small-sample permutation) produce conservative, non-uniform p -values by construction. Composite nulls with estimated nuisance parameters may exhibit slight departures even when properly calibrated.

VII. The Bayesian Alternative: Quantifying Support

Classical inference asks $P(\text{data} \mid H_0)$. Bayesian inference asks what practitioners actually want: $P(H_0 \mid \text{data})$.

Bayes’ theorem:

$$p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta) p(\theta).$$

Prior beliefs $p(\theta)$ are updated by the likelihood $p(\mathbf{y} \mid \theta)$ to yield the posterior $p(\theta \mid \mathbf{y})$.

The Bayes Factor. The primary Bayesian tool for model comparison:

$$\text{BF}_{10} = \frac{P(\mathbf{y} \mid H_1)}{P(\mathbf{y} \mid H_0)}.$$

Unlike the p -value, BF_{10} compares the likelihood of the data under *two competing models* simultaneously. $\text{BF}_{01} = 10$ means the data are 10 times more likely under H_0 than H_1 —providing direct, positive evidence *for* H_0 , not merely a failure to reject it.

The legal analogy resolved. If the frequentist is a juror deciding if there is “reasonable doubt” to convict, the Bayesian is a detective weighing relative evidence for two suspects. In the Bayesian framework, “not guilty” can actually mean “innocent” if evidence for that hypothesis is sufficiently overwhelming.

Practical equivalence (ROPE). When a parameter is never exactly zero, Bayesians use the **Region of Practical Equivalence (ROPE)** to define a range of negligible values. If the entire 95% Highest Density Interval (HDI) falls within the ROPE, the null is accepted for practical purposes—something a non-significant p -value cannot provide.

Feature	Frequentist	Bayesian
Conclusion	Cannot accept H_0	Can support H_0
Measure	Surprise under H_0	H_0 vs. H_1 ratio
Ambiguity	Large p uninformative	Distinguishes absence of evidence from evidence of absence

VIII. Effect Size and Practical Significance

Statistical significance and practical importance are orthogonal. With large n , any non-zero effect—however tiny—will eventually be statistically significant. With small n , a large and important effect may not reach significance.

Always report an effect size:

Context	Measure	Small/Med/Large
Means	Cohen’s d	0.2 / 0.5 / 0.8
ANOVA	η_p^2	0.01 / 0.06 / 0.14
Proportions	OR, RR, RD	context-dependent
Correlation	r	0.1 / 0.3 / 0.5
Regression	Cohen’s f^2	0.02 / 0.15 / 0.35

Here, OR-odds ratio, RR-relative risk, RD-risk difference.

Effect sizes are also essential for sample size planning: determine n *before* data collection, based on the minimum effect of practical interest, not post-hoc.

IX Confidence Intervals as Inference

A $100(1 - \alpha)\%$ CI is a random interval $[L, U]$ such that $P(L \leq \theta \leq U) = 1 - \alpha$ across repeated samples.

Correct interpretation: if this procedure were repeated many times, $1 - \alpha$ of the resulting intervals would contain θ . This is a statement about the *procedure*, not the specific interval computed from the current data.

CIs provide more than p -values: direction, magnitude, and precision of the estimate. If the entire CI lies within a region of trivial effects, the finding is practically negligible even if $p < 0.05$.

CI-test duality: a two-sided test rejects $H_0 : \theta = \theta_0$ at level α if and only if θ_0 falls outside the $(1 - \alpha)$ CI. The CI is the set of all null values not rejected.

Warning: a CI is not a Bayesian credible interval. They coincide numerically only under flat priors but carry fundamentally different interpretations.

X. Multiple Testing and the Forking Paths Problem

Multiple comparisons. When m tests are conducted simultaneously at level α , the probability of at least one false positive (family-wise error rate, FWER) is:

$$\text{FWER} = 1 - (1 - \alpha)^m.$$

For $m = 20$ and $\alpha = 0.05$: $\text{FWER} \approx 64\%$. Corrections—Bonferroni (α/m), Holm (step-down), Benjamini–Hochberg (FDR)—control different error rates.

The forking paths problem. A p -value is valid only when the analysis plan is fixed before seeing the data. Data-dependent choices—transformations, subsets, outlier removal, switching from one-sided to two-sided tests—constitute hidden multiplicity that inflates the true Type I error above the nominal level.

Two primary mechanisms:

1. **One-sided to two-sided (inflation):** switching after observing the direction of the result effectively doubles the rejection region, inflating α far above 0.05.

2. **Two-sided to one-sided (significance chasing):** if a two-sided test yields $p = 0.08$, switching to one-sided to report $p = 0.04$ uses the data to justify a more lenient test—a classic form of p -hacking.

The gold standard: the directional decision (one- vs. two-sided), the choice of test, and the analysis plan must be fixed *a priori*. If the direction of the effect is not known with certainty before the study begins, a two-sided test is the most robust and honest default.

Warning: post-hoc switching invalidates nominal significance levels. Integrity in analysis requires honoring the original hypothesis, regardless of the patterns that emerge in the data.

XI. Assumptions and Their Consequences

Every test in this compendium is valid under a set of assumptions. Violations do not necessarily invalidate the test, but they change its behavior.

Assumption	Consequence of violation
Independence	Biased SEs; inflated Type I error
Correct dist.	Non-standard null distribution
Equal variances	Inefficient; invalid SEs
No outliers	Inflated σ^2 ; masked effects
Correct model	Biased estimates

The appropriate response to an assumption violation depends on its severity:

- **Minor:** many tests are robust to mild violations (e.g. t -test under mild non-normality for large n).
- **Moderate:** use robust alternatives (Welch instead of pooled t ; HC SEs instead of OLS SEs).
- **Severe:** switch to a fundamentally different method (non-parametric, bootstrap, or a correctly specified model).

Bottom Line

- Statistical tests assess compatibility with models—not truth.
- Evidence is asymmetric in classical inference: tests detect departures, not confirmations.
- p -values measure compatibility with H_0 under assumptions—not the probability H_0 is true, not effect size, not importance.
- Effect sizes and confidence intervals are always required alongside p -values.

- Under H_0 , p -values are uniform on $[0, 1]$; deviations signal violated assumptions, data snooping, or unmodelled structure.
- Bayesian methods can quantify positive evidence for H_0 ; classical methods cannot.
- Pre-specification and transparency are not optional; they are the foundation of valid inference.
- Interpretation requires discipline beyond mechanical threshold crossing.

A Map of This Compendium

The tests in this book are organized by the type of inferential question they address. Every chapter primer provides the conceptual prerequisites for that chapter's tests.

Chapter	Core question/Context
NORM	Are the data normally distributed?
GFIT	Do the data fit a specified distribution?
MEANS	Do group means differ?
VARI	Do group variances differ?
PROP	Testing proportions and rates
NPAR	Rank-based alternatives to parametric tests
CORR	Are two variables associated?
CTAB	Are categorical variables independent?
REGR	Are regression assumptions satisfied?
GLMS	Does the GLM fit adequately?
MIXD	Mixed model: fixed and random effects
MULT	Multivariate mean and covariance tests
TS*	Time series: diagnostics (TSDG), stationarity (STAT), cointegration (COINT), volatility clustering (ARCH)
RAND	Randomness and independence
SURV	Time-to-event and censored data
SPAT	Spatial autocorrelation and regression
QUAL	Statistical process control
EQVL	Equivalence and non-inferiority
ROBU	Robust and resistant methods
PERM	Permutation and bootstrap inference
OUTL	Outlier detection
BAYES	Bayesian hypothesis testing and model comparison

Beneath their diverse formulas and assumptions, all statistical tests seek to quantify the tension between a model and the evidence.

PART I

Normality & Distributional Shape

Assessing distributional assumptions and shape characteristics

NORMALITY

I. NORM — A Conceptual Primer

1. Framing & Motivation

What Is Normality?

- Symmetric, bell-shaped distribution fully defined by (μ, σ^2) .
- In practice an **approximation**, never exact.
- Distinguish: **data normality** (raw observations) vs. **residual normality** (model errors)—only the latter is required for valid regression and ANOVA inference.

Why It Matters

- Required for **exact finite-sample** t -, F -, and χ^2 -inference.
- *Not* required for point estimation (OLS coefficients do not require normality; under the Gauss–Markov assumptions they remain BLUE).
- Large samples rely on the **CLT**; raw-data normality becomes progressively less critical.

Where Normality Is (and Is Not) Needed

Needed: Small-sample t -tests, ANOVA, regression inference, confidence intervals, prediction intervals.

Not needed: Point estimation, large- n inference, nonparametric methods, bootstrap-based inference.

2. Theoretical Foundation

The Normal Distribution

$$\text{Density: } f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Key properties:

- Mean = Median = Mode = μ .
- Skewness $\gamma_1 = 0$; excess kurtosis $\gamma_2 = 0$.
- **Empirical rule:** 68% within $\mu \pm \sigma$; 95% within $\mu \pm 2\sigma$; 99.7% within $\mu \pm 3\sigma$.
- Tails decay as $e^{-x^2/2}$ —much thinner than t or Cauchy.
- All odd central moments = 0; $E[(X-\mu)^4] = 3\sigma^4$ (benchmark for kurtosis).

Core Moment Formulas

$$\bar{x} = \frac{1}{n} \sum x_i, \quad m_k = \frac{1}{n} \sum (x_i - \bar{x})^k, \quad s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$
$$\gamma_1 = \frac{m_3}{m_2^{3/2}}, \quad \gamma_2 = \frac{m_4}{m_2^2} - 3 \quad (\text{excess kurtosis})$$

Useful Identities

- *Affine transform:* $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

- *Sum of independents:* $\sum X_i \sim N(\sum \mu_i, \sum \sigma_i^2)$.
- *Sample mean:* $\bar{X} \sim N(\mu, \sigma^2/n)$ exactly (or approx. via CLT).
- *Standardization:* $Z = (X - \mu)/\sigma \sim N(0, 1)$; $\Phi(-x) = 1 - \Phi(x)$. $\Phi(x)$ is the standard normal CDF.

Derived distributions (normality $\Rightarrow$):

$$\sum_{i=1}^k Z_i^2 \sim \chi_k^2, \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2, \quad \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

These exact results underpin t -, F -, and χ^2 -tests throughout the compendium.

CLT — Practitioner View

- The sampling distribution of $\bar{X}$ converges to $N(\mu, \sigma^2/n)$ as $n \rightarrow \infty$.
- Mild skew: $n \approx 30$ often sufficient.
- Heavy tails or outliers: larger n required.
- **CLT does not save you** in small samples or with extreme departures.

3. Structural Understanding of Normality

Characterizations of Normality

Three testable properties uniquely identify the normal distribution:

- **Moments:** $\gamma_1 = 0$ (symmetry) and excess kurtosis = 0 (mesokurtosis). *Caveat:* These conditions are necessary but not sufficient for normality; non-normal distributions can also have $\gamma_1 = 0$ and $\gamma_2 = 0$.
- **Quantile linearity:** order statistics $y_{(i)}$ are linear in expected normal scores $q_i = \Phi^{-1}((i - 0.375)/(n + 0.25))$ (Blom approximation).
- **CDF proximity:** $F_n(y)$ is uniformly close to $\Phi((y - \mu)/\sigma)$.

Skewness & Kurtosis — Shape Guide

- **Right skew** ($\gamma_1 > 0$): long right tail; mean > median; Q–Q bows below line at upper end.
- **Left skew** ($\gamma_1 < 0$): long left tail; mean < median; Q–Q bows above line at lower end.
- **Leptokurtosis** ($\gamma_2 > 0$): heavy tails, frequent extremes; Q–Q ends bend away (S-shape).
- **Platykurtosis** ($\gamma_2 < 0$): thin tails, few extremes; Q–Q ends bend toward the line.

4. Diagnostics

Q–Q Plot — Primary Diagnostic

Plot $y_{(i)}$ against q_i . A straight 45° line $\Rightarrow$ consistent with normality.

Pattern	Interpretation
Points on line	Normality
S-shape (ends out)	Heavy tails (leptokurtic)
Reverse S (ends in)	Thin tails (platykurtic)
Bow above line	Left skew
Bow below line	Right skew
Isolated extremes	Outliers

Focus on *patterns*, not minor random scatter. The Q–Q plot reveals *where* and *how* the distribution departs from normality—information formal tests cannot convey.

5. Testing Framework

Simple vs. Composite Hypothesis

- **Simple H_0 :** μ_0 and σ_0 known; standard K–S critical values apply.
- **Composite H_0 :** μ and σ estimated from data (almost always the practical case); critical values must be adjusted. Using simple-hypothesis tables is **anti-conservative**. The Lilliefors correction is the canonical fix for K–S.

Taxonomy of Tests

Class	Based on	Examples
Regression	Q–Q linearity	Shapiro–Wilk, Shapiro–Francia
EDF	F_n vs. Φ distance	K–S, Lilliefors, Anderson–Darling, Cramér–von Mises
Moment	γ_1, γ_2	Jarque–Bera, D’Agostino–Pearson

Each class is sharpest against the departures it directly targets: moment tests detect skewness/kurtosis; EDF tests detect general distributional shifts; regression tests are sensitive to overall Q–Q non-linearity. **Shapiro–Wilk** has the best omnibus power across a wide range of alternatives and is the general-purpose default.

Power & Sample Size

- **Small n (< 30):** low power—tests most needed are least informative. Rely on Q–Q plot; use Shapiro–Wilk.
- **Large n (> 200):** very high power—trivial, practically irrelevant departures are flagged. CLT provides protection for downstream tests regardless.
- A non-significant result **does not confirm normality**; it only fails to reject it.

6. Consequences & Decision Layer

Impact of Violations

- **Mild skew:** usually acceptable, especially with balanced designs.
- **Heavy tails:** inflate Type I error; t -test is sensitive.
- **Outliers:** most damaging; distort both estimates and tests.
- **Multimodality:** signals model misspecification, not just non-normality.

When You Can Safely Ignore Normality

- Large sample size (CLT applies).
- No extreme outliers present.
- Mild departures only (modest skew, moderate kurtosis).
- Robust or bootstrapped standard errors are used.

7. Actionable Remedies

What To Do If Normality Fails

- **Transform:** $\log y$ (right skew), $\sqrt{y}$ (counts), Box–Cox λ (data-driven).
- **Nonparametric methods:** Wilcoxon, Kruskal–Wallis (see NPAR chapter).
- **Bootstrap inference:** no distributional assumption needed.
- **Robust methods:** trimmed means, M -estimators (see ROBU chapter).
- **Re-specify model:** non-normality often signals omitted variables or wrong functional form—the root cause, not a symptom.

Robust Measures (Normality-Free)

- Median, $IQR = Q_3 - Q_1$.
- $MAD = \text{median}(|x_i - \text{median}(\mathbf{x})|)$; multiply by 1.4826 for σ -consistent estimate under normality.
- Less sensitive to outliers and heavy tails than $\bar{x}$ and s .

8. Practitioner Guardrails & Closure

Common Pitfall

Testing normality of *raw predictors* in regression.

Fix: Always assess **residuals** $\hat{\varepsilon}_i$, not the original data.

Bottom Line

Normality is a tool for inference, not a requirement to enforce. Most real data are non-normal—and that is often acceptable. Test *residuals*, not raw data. Combine Q–Q plots with formal tests. At large n , the CLT is your friend.

Shapiro-Wilk Test

Powerful Test for Assessing Normality

NORM.1 | **Category:** Normality | **Data:** Continuous | **Null:** Normal

When to Use

- Assess normality of a single continuous variable.
- Routine diagnostic prior to t -tests, ANOVA, and linear regression.
- Effective for sample sizes: $3 \leq n \leq 5000$.

Assumptions

- Random sample with i.i.d observations.
- Continuous data with few or no ties.
- Univariate measurement (one variable at a time).
- Complete, uncensored observations.

Test Statistic

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \left| \quad \mathbf{a} = \frac{\mathbf{V}^{-1} \mathbf{m}}{\|\mathbf{V}^{-1} \mathbf{m}\|}, \quad m_i = \Phi^{-1} \left(\frac{i-a}{n+b} \right) \quad \left| \quad 0 \text{ (non-normal)} < W \leq 1 \text{ (normal)} \right.$$

Notation: $x_{(i)}$ are ordered sample values, $m_i = E[X_{(i)}]$, $V_{ij} = \text{Cov}(X_{(i)}, X_{(j)})$, Blom's approx: $a = 3/8$, $b = 1/4$

Plain English: The Shapiro-Wilk test measures how well ordered data align with what would be expected under normality. It quantifies the linearity of a normal Q-Q plot by comparing two variance estimates (R^2 like metric): one from regression of ordered values (numerator) on expected normal order statistics and another from sample variance (denominator).

⚠ Practitioner Warning

The test becomes highly sensitive in large samples ($n > 1000$); even trivial deviations from normality may yield significant results. Always complement with graphical diagnostics (Q-Q plots, histograms) and consider practical significance.

Hypotheses

- H_0 : Data distribution is Normal
- H_1 : Data distribution is **not** Normal

Decision Rule: Reject H_0 if $W < W_{\alpha,n}$, where $W_{\alpha,n}$ is the critical value for significance level α and sample size n .

📦 Software References [Python, R, SAS]

`scipy.stats.shapiro`

`shapiro.test`

PROC UNIVARIATE / NORMAL

Additional Notes

The exact distribution of W under H_0 is complex and depends on n . Critical values for $\alpha = 0.01, 0.05, 0.10$ have been tabulated for $3 \leq n \leq 50$. For $n > 50$, Royston's (1992) approximations ($w \sim N(\mu, \sigma^2)$) are used. The test has excellent power against most alternatives to normality.

$4 \leq n \leq 11$

$$w = -\ln[\gamma - \ln(1 - W)]; \quad \gamma = -2.273 + 0.459n$$

$$\mu = 0.5440 - 0.39978n + 0.025054n^2 - 0.0006714n^3$$

$$\sigma = \exp(1.3822 - 0.77857n + 0.062767n^2 - 0.0020322n^3)$$

$12 \leq n \leq 2000$

$$w = \ln(1 - W); \quad x = \ln n$$

$$\mu = -1.5861 - 0.31082x - 0.083751x^2 + 0.0038915x^3$$

$$\sigma = \exp(-0.4803 - 0.082676x + 0.0030302x^2)$$

Prior normality tests (such as the χ^2 goodness-of-fit) were less powerful or required grouping data into bins, which lost information. Shapiro and Wilk's insight was to leverage the properties of order statistics to create a test that used the full dataset without binning. The test quickly became the gold standard due to its superior power. However, computing W by hand was tedious—the coefficients a_i required inverting covariance matrices of order statistics. Early adoption was slow until Royston (1982) provided efficient algorithms and extended the method to larger sample sizes. *Note:* A censored-data version of the test is available for assessing normality in partially observed samples.

Miscellanea: The original paper is titled “An Analysis of Variance Test for Normality”—a subtle misnomer, as the test is not an ANOVA in the usual sense. Interestingly, Wilk also co-invented (with R. Gnanadesikan, 1968) the Q-Q plot, so the test and its most widely used diagnostic share a common origin. *Mnemonic:* “W for Wilk, W for Well-behaved”—if $W \approx 1$, data are behaving *normally*.

Reference(s): Shapiro & Wilk (1965); Royston (1982, 1992)

Worked Example

Problem Statement Eight recovery times (days) are tested for normality before applying a parametric procedure:

Sorted Data and Sample Moments Sorted: $x_{(1)} \leq \dots \leq x_{(8)}$:

$$3, 5, 6, 7, 9, 12, 18, 35 \quad \bar{x} = 11.875, \quad SS = \sum_{i=1}^8 (x_i - \bar{x})^2 = 764.88$$

H_0 : the data are drawn from a normal distribution; H_1 : the data are not normally distributed.

1. Shapiro-Wilk Coefficients & Numerator b

For $n = 8$, the tabulated coefficients a_i are applied to symmetric order-statistic differences $d_i = x_{(n+1-i)} - x_{(i)}$:

i	a_i	$x_{(9-i)}$	$x_{(i)}$	d_i
1	0.6052	35	3	32
2	0.3164	18	5	13
3	0.1743	12	6	6
4	0.0561	9	7	2

$$b = \sum_{i=1}^4 a_i d_i = 24.638$$

2. W Statistic

$$W = \frac{b^2}{SS} = \frac{(24.638)^2}{764.875} = \frac{607.011}{764.875} = 0.7936$$

$W \in (0, 1]$; a value near 1 is consistent with normality. Here $W = 0.794$ indicates substantial departure.

3. Results and Decision

Statistic	Value	CV (5%)	Decision
W	0.7936	0.818	Reject H_0

$p = 0.024$; $W = 0.794 < W_{\text{crit}}(n=8, 5\%) = 0.818$. For reference: $W_{\text{crit}}(n=8, 10\%) = 0.851$.

4. Comparison of SS contributions

x_i	3	5	6	7	9	12	18	35	$\sum (x_i - \bar{x})^2$
$(x_i - \bar{x})^2$	78.77	47.27	34.52	23.77	8.27	0.02	37.52	534.77	764.88

5. Interpretation

- Normality is rejected at 5%. The value $x = 35$ lies 2.21 standard deviations above the mean and inflates SS from ≈ 230 (without it) to 764.9, while the numerator b grows only moderately because $a_1 = 0.605$ is bounded. The resulting $W = 0.794$ signals right-skewed departure (sample skewness = 1.50).

Practitioner Insight: The Shapiro–Wilk statistic assesses how closely ordered data match expected normal order statistics. The coefficients a_i reflect normal spacings, so $W = b^2/SS$ is essentially the squared correlation in a normal probability plot, with $W = 1$ under perfect normality. It is sensitive to skewness and heavy tails, and is generally the most powerful normality test for $n \leq 50$. For larger samples, Royston’s extension provides analytic p -values. Rejection indicates non-normality but not its form, so graphical tools (e.g. Q–Q plots) are needed to diagnose departures.

Common Pitfalls

- Non-rejection as confirmation.** Failing to reject H_0 indicates insufficient evidence *against* normality, not proof that the data are normal. Always complement the test with Q–Q plots and histograms.
- Large-sample over-sensitivity.** With large n , even trivial deviations from normality can produce highly significant results. Assess whether departures are practically meaningful for the intended analysis.
- Very small samples.** For very small samples ($n < 10$), the test has limited power and may fail to detect substantial non-normality. Results should be interpreted cautiously.
- Tied observations.** Shapiro–Wilk assumes continuous data. Extensive ties—often due to rounding or discrete measurement—can distort the test statistic and invalidate inference.
- Inherently non-normal data.** Counts, proportions, and bounded variables are not normally distributed. Normality testing is inappropriate; use models suited to the data type instead.

Shapiro-Francia Test

Simplified Approximation to Shapiro-Wilk for Normality

NORM.2 | **Category:** Normality | **Data:** Continuous | **Null:** Normal

When to Use

- Designed as a large-sample approximation to Shapiro–Wilk.
- Particularly suitable for moderate to large sample sizes.
- Useful when computational simplicity is desired.

Assumptions

- Random sample with i.i.d observations.
- Continuous data with few or no ties.
- Complete, uncensored observations.

Test Statistic

$$W' = \frac{(\sum_{i=1}^n b_i x_{(i)})^2}{\sum_{i=1}^n (x_{(i)} - \bar{x})^2} \quad \Bigg| \quad \mathbf{b} = \frac{\mathbf{m}}{\|\mathbf{m}\|} = \frac{m_i}{\sum_{i=1}^n m_i^2}, \quad m_i = \Phi^{-1}\left(\frac{i-a}{n+b}\right) \quad \Bigg| \quad 0 \text{ (non-normal)} < W' \leq 1 \text{ (normal)}$$

Notation: $x_{(i)}$ are order statistics, $m_i = E[X_{(i)}]$, Blom's approx: $a = 3/8$, $b = 1/4$, Φ^{-1} is inv. standard normal CDF

Plain English: The Shapiro-Francia test measures how strongly the ordered data correlate with expected values from a normal distribution. It simplifies the Shapiro-Wilk calculation by using expected normal order statistics directly rather than computing inverse covariance matrices, making it more practical for larger samples.

⚠ Practitioner Warning

The Shapiro–Francia test is designed primarily to detect departures from normality through tail behaviour and overall distributional shape. While it performs well for large samples and approximately symmetric alternatives, it may be less sensitive than the Shapiro–Wilk test to certain subtle deviations from normality. As with all normality tests, statistical significance should be interpreted alongside graphical diagnostics such as Q–Q plots, since large samples can detect negligible departures and small samples may lack power to detect meaningful ones.

Hypotheses

- H_0 : Data distribution is Normal
- H_1 : Data distribution is **not** Normal

Decision Rule: Reject H_0 if $W' < W'_{\alpha,n}$, where $W'_{\alpha,n}$ is the critical value for significance level α and sample size n .

📦 Software References [Python, R, SAS]

Not Native (Manual)

`nortest::sf.test`

Not Native (Manual)

Additional Notes

The Shapiro–Francia statistic can be interpreted geometrically as the squared Pearson correlation coefficient between the ordered observations and their expected normal scores:

$$W' = \frac{(\sum_{i=1}^n (x_{(i)} - \bar{x})(m_i - \bar{m}))^2}{\sum_{i=1}^n (x_{(i)} - \bar{x})^2 \sum_{i=1}^n (m_i - \bar{m})^2}, \quad \text{where } \bar{m} = \frac{1}{n} \sum_{i=1}^n m_i.$$

Francia's insight was that expected normal order statistics m_i could serve as reasonable approximations to the Shapiro-Wilk's optimal weights (a_i). While this sacrificed some statistical power, it made the test calculable with just a table of normal quantiles and a hand calculator. Royston (1993) showed that, under normality and for moderate to large sample sizes,

$$A = \ln(1 - W') \approx \mathcal{N}(\mu_n, \sigma_n^2), \quad \text{where } \mu_n = -1.2725 + 1.0521 [\ln(\ln n) - \ln n], \quad \sigma_n = 1.0308 - 0.26758 [\ln(\ln n) + \frac{2}{\ln n}].$$

Miscellanea: Francia introduced this test in 1972 as a “poor man’s Shapiro–Wilk”—a computationally simpler alternative at a time when computing resources were limited, just seven years after the original Shapiro–Wilk test. Ironically, modern software now makes both tests equally easy to run. The method is also closely related to earlier work by de Wet and Venter, who studied correlation-type statistics based on normal order statistics; all these approaches compare ordered data to their expected normal counterparts, reflecting a broader regression-based view of normality testing.

Reference(s): Shapiro & Francia (1972); Royston (1983, 1993)

Worked Example

Problem Statement Ten ($n = 10$) hospital length-of-stay values (days) are tested for normality using the Shapiro-Francia test. Unlike Shapiro-Wilk, which derives weights from the full covariance matrix of normal order statistics, Shapiro-Francia uses expected normal scores directly. H_0 : the data are drawn from a normal distribution; H_1 : they are not.

1. Normal Scores via Blom Approximation

For order statistic $x_{(i)}$, the expected standard-normal quantile is:

$$m_i = \Phi^{-1}(p_i), \quad \text{where } p_i = \frac{i - 0.375}{n + 0.25}$$

i	p_i	m_i	$x_{(i)}$	$m_i x_{(i)}$
1	0.0610	-1.5466	5	-7.733
2	0.1585	-1.0005	7	-7.003
3	0.2561	-0.6554	8	-5.243
4	0.3537	-0.3755	9	-3.379
5	0.4512	-0.1226	10	-1.226
6	0.5488	+0.1226	12	+1.471
7	0.6463	+0.3755	15	+5.632
8	0.7439	+0.6554	20	+13.109
9	0.8415	+1.0005	28	+28.014
10	0.9390	+1.5466	52	+80.425
$\sum m_i x_{(i)}$				104.065

Note: $\bar{m} = 0$ exactly by the symmetry of the Blom formula, so $S_{mx} = \sum m_i x_{(i)} - n\bar{m}\bar{x} = \sum m_i x_{(i)}$.

5. Interpretation

- $W' = 0.748$ falls below all three critical values. The dominant signal is the outlying value $x = 52$ ($m_{10}x_{(10)} = 80.4$, accounting for 77% of S_{mx}), which inflates SS_{xx} far more than it inflates S_{mx} , pulling r below 1.

Practitioner Insight: W' is the squared correlation between ordered data $x_{(i)}$ and Blom normal scores m_i , so $W' = 1$ under perfect normality and decreases with curvature (skewness, heavy tails, outliers) in the $Q-Q$ plot. Using m_i simplifies computation relative to Shapiro-Wilk, though it is slightly less powerful for small $n \leq 11$. For $n > 50$, both tests perform similarly and W' is preferred for simplicity; p -values are obtained via the Royston approximation based on $\log(1 - W')$.

2. Sample Moments and Score Variance

$$\bar{x} = 16.6, \quad SS_{xx} = \sum (x_i - \bar{x})^2 = 1820.4$$

$$SS_{mm} = \sum m_i^2 = 7.957 \quad (\bar{m} = 0)$$

3. $Q-Q$ Correlation and W'

$$r = \frac{S_{mx}}{\sqrt{SS_{mm} \cdot SS_{xx}}} = \frac{104.065}{\sqrt{7.957 \times 1820.4}} = 0.8646$$

$$W' = r^2 = (0.8646)^2 = 0.7476$$

W' is the R^2 between the order statistics and their expected normal positions — equivalently, the R^2 of the $Q-Q$ plot regression line through the origin.

4. Results and Decision

Statistic	Value	CV (1%)	Decision
W'	0.7476	0.793	Reject H_0

α	$W'_{\text{crit}}(n=10)$	Decision
10%	0.869	Reject
5%	0.842	Reject

Common Pitfalls

- Small-sample limitations.** Shapiro-Francia performs relatively poor for $n < 20$; use Shapiro-Wilk in this range due to markedly higher power.
- Statistic mismatch.** The W' statistic is not interchangeable with Shapiro-Wilk's W ; they have distinct null distributions and critical values.
- Simplicity $\neq$ superiority.** Shapiro-Francia was created for computational ease; modern software makes Shapiro-Wilk just as easy and generally more powerful.
- Overreliance on the test.** As with all normality tests, important structure may appear only in $Q-Q$ plots or histograms; visual checks remain essential.
- Inappropriate data types.** The test assumes continuous, unbounded data; discrete, rounded, or bounded variables (counts, proportions, Likert scales) violate this assumption and require alternative methods.

Anderson–Darling Test

Tail-Sensitive Test for Normality

NORM.3 | **Category:** Normality | **Data:** Continuous | **Null:** Normal

When to Use

- Test normality with emphasis on detecting deviations in distribution tails.
- Suitable for a wide range of sample sizes ($n \geq 5$).
- More powerful than Kolmogorov-Smirnov for detecting tail departures.

Assumptions

- Random sample with i.i.d observations.
- Continuous data with few or no ties.
- Complete, uncensored observations.
- When parameters are estimated, use modified version of the test (Anderson-Darling with estimated parameters).

Test Statistic

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n [(2i-1) \ln(\Phi(z_{(i)})) + (2n+1-2i) \ln(1-\Phi(z_{(i)}))] \quad \Bigg| \quad (\text{Stephens}): A^{*2} = A_n^2 \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2} \right)$$

Notation: $z_{(i)}$ are the *standardized* order statistics, $\Phi(\cdot)$ is the standard normal CDF, A_n^{2*} corrects for finite sample.

Plain English: It measures the weighted discrepancy between the empirical distribution and a normal distribution, giving extra weight to observations in the tails. This makes it particularly sensitive to departures that occur in the extremes of the distribution—exactly where violations often matter most.

⚠ Practitioner Warning

When using sample mean and variance to define the normal distribution, standard critical values are invalid. Use the modified Anderson-Darling test with adjusted critical values or directly obtain p -values from software.

Hypotheses

- H_0 : Data distribution is Normal
- H_1 : Data distribution is **not** Normal

Decision Rule: Reject H_0 if $A^{*2} > A_{\alpha}^{*2}$ (**critical value**)
or $p\text{-value} < \alpha$. $A_{0.01}^{*2} = 1.035$, $A_{0.05}^{*2} = 0.752$, $A_{0.10}^{*2} = 0.631$

📦 Software References [Python, R, SAS]

`scipy.stats.anderson`

`nortest::ad.test`

PROC UNIVARIATE / NORMAL

Additional Notes

Anderson and Darling (1952) introduced the test as a modification of the Cramér-von Mises statistic, incorporating a weighting function that produces consistent tests against all alternatives. The Anderson-Darling statistic (definition) for testing whether F_n (empirical CDF) comes from a continuous distribution F is:

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{F(x)[1 - F(x)]} dF(x).$$

The weighting function $w(x) = [F(x)(1 - F(x))]^{-1}$ gives more weight to observations in the tails, where $F(x)$ is close to 0 or 1. The test works for any general continuous distribution F . Subsequent work by Stephens (1974) provided critical values and modifications for parameters estimated from data, making the test practical for routine application.

Unlike the Kolmogorov-Smirnov test, which focuses on the maximum discrepancy between empirical and theoretical distributions, the Anderson-Darling test integrates squared discrepancies across the entire distribution, weighted to emphasize tail behavior. This weighting makes it one of the most powerful tests for normality, especially against alternatives with non-normal tail behavior.

Miscellanea: The test remained difficult to use in practice until 1974, when Michael A. Stephens derived modified critical values that made it distribution-free, transforming it from a theoretical curiosity into the widely-used test found in statistical software today. Stephens' modification was so successful that most modern implementations labeled "Anderson-Darling test" actually use his 1974 version. Unlike Shapiro-Wilk, Anderson-Darling is a general goodness of fit test (not just *normal*).

Reference(s): Anderson & Darling (1952); Stephens (1986); Marsaglia & Marsaglia (2004)

Worked Example

Problem Statement Eight service times (minutes) are tested for normality. The Anderson-Darling test is chosen over Kolmogorov-Smirnov because it weights deviations in the tails of the distribution more heavily than the center. H_0 : data follow a normal distribution; H_1 : they do not.

1. Standardize the Order Statistics $\bar{x} = 18.75$, $s = 8.972$, $z_{(i)} = (x_{(i)} - \bar{x})/s$.

i	$x_{(i)}$	$z_{(i)}$	$\Phi(z_{(i)})$	$1 - \Phi_{n+1-i}$	$2i - 1$	$\log \Phi(z_{(i)})$	$\log(1 - \Phi_{n+1-i})$
1	10	-0.9752	0.1647	.0120	1	-1.804	-4.423
2	13	-0.6409	0.2608	.3586	3	-1.344	-1.026
3	14	-0.5294	0.2983	.4889	5	-1.210	-0.716
4	16	-0.3065	0.3796	.5773	7	-0.969	-0.549
5	17	-0.1950	0.4227	.6204	9	-0.861	-0.477
6	19	+0.0279	0.5111	.7017	11	-0.671	-0.354
7	22	+0.3622	0.6414	.7392	13	-0.444	-0.302
8	39	+2.2570	0.9880	.8353	15	-0.012	-0.180

2. AD Statistic (Alternative Formula)

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i - 1) [\log \Phi(z_{(i)}) + \log(1 - \Phi(z_{(n+1-i)}))] = -8 - \frac{1}{8}(-69.496) = \mathbf{0.687}$$

3. Modified Statistic and Decision

Adjusting for estimated $\bar{x}$ and s :

$$A^{*2} = A_n^2 \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2} \right) = 0.687 \times 1.129 = \boxed{0.776}$$

Statistic	Value	CV (5%)	Decision
A^{*2}	0.776	0.752	Reject H_0

$p \approx 0.044$. Reject H_0 at 5%.

Practitioner Insight: The Anderson–Darling statistic weights observations by $(2i - 1)$, placing greatest emphasis on the tails. Large contributions arise when tail probabilities are extreme, with asymmetry between Φ_i and $1 - \Phi_{n+1-i}$ revealing skewness. Unlike Cramér–von Mises, which weights all points equally, AD is more sensitive to tail deviations. The adjustment $A^{*2} = A_n^2(1 + 0.75/n + 2.25/n^2)$ corrects for estimating μ and σ . For large samples, even minor deviations can trigger rejection, so graphical checks (e.g. Q–Q plots) are essential; here, a single extreme value drives the result.

Common Pitfalls

- **Unadjusted critical values.** When μ and σ are estimated from the data, the raw A_n^2 statistic is invalid. Always use the adjusted statistic A^{*2} or software that applies the correction automatically.
- **Ties and discretization.** The Anderson–Darling test assumes continuous data. Extensive ties—often caused by rounding or discrete scales—can distort the statistic and invalidate inference.
- **p -value only interpretation.** The test indicates whether normality fails, not how. Always inspect Q–Q plots to identify whether deviations occur in the tails, center, or via skewness.
- **Large-sample over-sensitivity.** With large n , even negligible departures from normality can be statistically significant. Assess whether the deviation is practically meaningful.
- **Assumed dominance.** Anderson–Darling is especially sensitive to tail departures but is not uniformly most powerful. Tests such as Shapiro–Wilk may outperform it for other alternatives.

 p -value Formulas

For $A^{*2} \geq 0.600$, $p = \exp(1.2937 - 5.709A^{*2} + 0.0186(A^{*2})^2)$; for $0.340 < A^{*2} < 0.600$, $p = \exp(0.9177 - 4.279A^{*2} - 1.38(A^{*2})^2)$; for $0.200 < A^{*2} \leq 0.340$, $p = 1 - \exp(-8.318 + 42.796A^{*2} - 59.938(A^{*2})^2)$; and for $A^{*2} \leq 0.200$, $p = 1 - \exp(-13.436 + 101.14A^{*2} - 223.73(A^{*2})^2)$.

Cramér–von Mises Test

Integrated Squared Distance Goodness-of-Fit Test

NORM.4 | **Category:** Normality | **Data:** Continuous | **Null:** Normal

When to Use

- Test normality with balanced sensitivity across the entire distribution.
- Preferred when no specific region (center vs. tails) is of primary concern.
- More powerful than Kolmogorov–Smirnov, simpler than Anderson–Darling.

Assumptions

- Random sample with i.i.d observations.
- Continuous data with few or no ties.
- Complete, uncensored observations.
- When parameters are estimated, use modified critical values.

Test Statistic

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^n \left[\frac{2i-1}{2n} - \Phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right) \right]^2 \quad (\text{Stephens}): W^{*2} = W_n^2 \left(1 + \frac{0.5}{n}\right)$$

This form is specific to testing normality, with (μ, σ^2) estimated from the data. Cramér–von Mises critical values are sample-size-free under a fully specified null (say, known μ_0, σ_0^2), but depend on it when (μ, σ^2) are estimated.

Notation: $x_{(i)}$ denotes the i th ordered observation, $\hat{\mu}$ and $\hat{\sigma}$ are the estimated sample mean and standard deviation, and $\Phi(\cdot)$ is the standard normal CDF. **Note:** Shapiro–Wilk uses W while Cramér–von Mises uses W^2 (for squared distance)

Plain English: The Cramér–von Mises test measures the integrated squared difference between the empirical and theoretical distributions, treating all regions equally. Unlike Anderson–Darling which emphasizes tails, this test is balanced—equally sensitive to departures anywhere in the distribution.

⚠ Practitioner Warning

By treating all regions equally, Cramér–von Mises sacrifices some power against specific alternatives (e.g., heavy tails) compared to targeted tests. Choose based on whether you care about all departures equally or have specific concerns about tails or center.

Hypotheses

- H_0 : Data distribution is Normal
- H_1 : Data distribution is **not** Normal

Decision Rule: Reject H_0 if $W^2 > W_{\alpha,n}^2$ (critical value) or $p\text{-value} < \alpha$. $W_{0.01}^2 = 0.179$, $W_{0.05}^2 = 0.126$, $W_{0.10}^2 = 0.104$.

📦 Software References [Python, R, SAS]

`scipy.stats.cramervonmises`

`nortest::cvm.test`

PROC UNIVARIATE / NORMAL

Additional Notes

The Cramér–von Mises test, independently introduced by Cramér (1928) and von Mises (1931), measures goodness of fit by integrating squared differences between the empirical and theoretical distribution functions (generic version),

$$W_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 dF(x),$$

providing a smooth, omnibus alternative to supremum-based tests such as Kolmogorov–Smirnov. Under H_0 , W_n^2 converges to a weighted sum of independent χ_1^2 variables, with critical values obtained from the asymptotic distribution or finite-sample tables.

Miscellanea: The Cramér–von Mises test is a classic example of scientific synchronicity. Harald Cramér (1928) and Richard von Mises (1931) independently derived essentially the same goodness-of-fit statistic while working in different countries and fields; their contributions were later unified under a single test. *Mnemonic:* CvM measures the *cumulative* (integrated) gap between distributions.

Reference(s): Cramér (1928); von Mises (1931); Anderson & Darling (1952)

Worked Example

Problem Statement Eight reaction times (milliseconds) are tested for normality using the Cramér–von Mises test. Unlike Anderson–Darling (which up-weights the tails) and Kolmogorov–Smirnov (which uses only the largest pointwise gap), CvM accumulates squared discrepancies between the empirical and fitted CDFs uniformly across all order statistics.

H_0 : data follow a normal distribution; H_1 : they do not.

1. Sample Moments and Standardization $\bar{x} = 12.5$, $s = 11.723$, $z_{(i)} = (x_{(i)} - \bar{x})/s$

2. CvM Statistic

For each order statistic, compute the gap between the fitted normal CDF $F_i = \Phi(z_{(i)})$ and the empirical midpoint $p_i = (2i - 1)/2n = (2i - 1)/16$:

i	$x_{(i)}$	$z_{(i)}$	F_i	p_i	$(p_i - F_i)^2$
1	4	-0.7251	0.2342	1/16	0.02948
2	5	-0.6398	0.2612	3/16	0.00543
3	7	-0.4692	0.3195	5/16	0.00005
4	8	-0.3839	0.3505	7/16	0.00756
5	9	-0.2986	0.3826	9/16	0.03235
6	11	-0.1280	0.4491	11/16	0.05684
7	16	+0.2986	0.6174	13/16	0.03808
8	40	+2.3458	0.9905	15/16	0.00281
$\sum (p_i - F_i)^2$					0.17260

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^n (p_i - F_i)^2 = \frac{1}{96} + 0.17260 = \mathbf{0.1830}$$

3. Modified Statistic

Adjusting for estimated $\bar{x}$ and s :

$$W^{*2} = W_n^2 \left(1 + \frac{0.5}{n} \right) = \boxed{0.1945}$$

4. Results and Decision

Statistic	Value	CV (1%)	Decision
W^{*2}	0.1945	0.179	Reject H_0

Full critical value table (Stephens 1974):

α	CV	W^{*2}	Decision
10%	0.104	0.1945	Reject
5%	0.126	0.1945	Reject
1%	0.179	0.1945	Reject

$p < 0.01$. Normality is rejected at all standard levels.

5. Interpretation

- The dominant terms are $i = 6$ ($x = 11$, gap = -0.238), $i = 7$ ($x = 16$, gap = -0.195), and $i = 5$ ($x = 9$, gap = -0.180). The extreme value $x = 40$ contributes only 0.003 — the body of the distribution, not the tail, drives the rejection.

Practitioner Insight: The CvM statistic W_n^2 approximates $n \int [F_n(x) - F(x)]^2 dF(x)$ by averaging squared gaps $(p_i - F_i)^2$ at the empirical CDF midpoints, giving equal weight to all observations. Unlike Anderson–Darling, which emphasizes tails, CvM is most sensitive to deviations in the body of the distribution; here, mid-range observations dominate while the extreme value contributes little. The adjustment $W^{*2} = W_n^2(1 + 0.5/n)$ corrects for estimating μ and σ , and Stephens' critical values (valid for $n \geq 5$) apply via this normalization.

Common Pitfalls

- Using incorrect critical values when parameters are estimated.** Standard tables assume a fully specified null, so when parameters (e.g. μ, σ) are estimated, modified CVs are required; using unadjusted values inflates Type I error.
- Expecting the test to be uniformly most powerful against all alternatives.** The equal-weighting nature of W^2 provides balanced sensitivity across the distribution, but it is often less powerful than specialized tests such as Anderson–Darling for tail deviations or Shapiro–Wilk for overall departures.
- Applying the test to data with many ties or discrete structure.** The test assumes continuous data, and ties from rounding or discreteness can distort the null distribution, leading to conservative or anti-conservative results.
- Relying solely on the test statistic without graphical diagnostics.** The W^2 statistic summarizes overall fit but does not indicate the nature of deviations, so Q–Q plots, histograms, and density estimates should also be examined.
- Interpreting non-rejection as proof of normality.** Failing to reject H_0 indicates only that the data are consistent with the model, and limited power—especially in small samples—means meaningful departures may go undetected.

Kolmogorov–Smirnov Test (Lilliefors Test)

Distribution-Free Test Based on Maximum Deviation

NORM.5

Category: Normality

Data: Continuous

Null: Distribution-free

When to Use

- Compare an empirical distribution to a fully specified theoretical distribution (not just normal distribution).
- Simple, distribution-free test requiring minimal assumptions.
- Suitable for any sample size, though power is moderate.
- Applicable beyond normality testing.

Assumptions

- Observations are independent and identically distributed.
- Data are continuous (test is conservative with ties).
- Null distribution must be **fully specified** (no parameter estimation).
- For normality with estimated parameters, use Lilliefors modification.

Test Statistic

$$D = \sup_x |F_n(x) - F_0(x)| = \max \left\{ \max_i \left| \frac{i}{n} - F_0(x_{(i)}) \right|, \max_i \left| F_0(x_{(i)}) - \frac{i-1}{n} \right| \right\}$$

Notation: $F_n(x)$ is empirical CDF, $F_0(x)$ is fully specified null CDF ($\Phi(x)$ for normal), supremum is maximum

Plain English: The Kolmogorov-Smirnov test measures the largest vertical gap between the empirical distribution and the hypothesized distribution. It asks: “What’s the worst mismatch anywhere?” Unlike tests using squared distances, K-S focuses on the single point of greatest disagreement.

⚠ Practitioner Warning

The standard K-S test **cannot** be used when parameters are estimated from data. Doing so severely inflates Type I error. For normality testing with estimated μ and σ , use the Lilliefors correction instead, which adjusts the null distribution appropriately.

Hypotheses

- H_0 : The data follow the specified distribution.
- H_1 : The data do not follow the specified distribution.

Decision Rule: Reject H_0 if $D > D_{\alpha,n}$ (critical value) or $p\text{-value} < \alpha$. For large n , $D_{\alpha,n} \approx c_\alpha / \sqrt{n}$ where $c_{0.01} \approx 1.63$, $c_{0.05} \approx 1.36$, $c_{0.10} \approx 1.22$.

📦 Software References [Python, R, SAS]

statsmodels.stats.
diagnostic.lilliefors

nortest::lillie.test

PROC UNIVARIATE / NORMAL

Additional Notes

Introduced by Kolmogorov (1933) and refined by Smirnov (1948), the Kolmogorov–Smirnov (K–S) test provides a distribution-free procedure for comparing an empirical distribution to a fully specified theoretical distribution, requiring minimal assumptions and making it broadly applicable. Finite-sample accuracy can be improved using Miller’s scaling correction (μ and σ^2 known),

$$D_n^{(M)} = \left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}} \right) D_n, \quad p = P \left(\sup_{0 \leq t \leq 1} |B(t)| \geq \sqrt{n} D_n \right) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp \left(-2k^2 (\sqrt{n} D_n)^2 \right),$$

where p denotes the p-value computed from the asymptotic Kolmogorov distribution of the Brownian bridge $B(t)$. The scaling reduces the liberal behavior of the unadjusted test in small and moderate samples.

Miscellanea: Andrey Kolmogorov—who established the axiomatic foundations of probability theory (often called the Newton of probability theory) in 1933—derived the asymptotic distribution of the K–S test that same year as the supremum of a **Brownian bridge** (a continuous stochastic process constrained to start and end at zero). This limiting distribution is now known as the **Kolmogorov distribution**. Despite the joint name, they never co-authored a paper on the test. *Mnemonic:* “K–S: King of Supremum”.

Reference(s): Kolmogorov (1933); Smirnov (1939, 1948); Miller (1951); Lilliefors (1967); Abdi & Molin (2007)

Worked Example

Problem Statement Eight pollution concentration measurements ($\mu\text{g}/\text{m}^3$) are tested for normality. When μ and σ are estimated from the data, the standard KS critical values are too conservative; the Lilliefors (1967) correction must be used. H_0 : data follow a normal distribution; H_1 : they do not.

1. Sample Moments and Standardization $\bar{x} = 12.0$, $s = 12.773$, $z_{(i)} = (x_{(i)} - \bar{x})/s$

2. KS Statistic: Both One-Sided Gaps

At each order statistic $x_{(i)}$, the empirical CDF jumps from $(i-1)/n$ to i/n . Two gaps are computed:

$$D_i^+ = \frac{i}{n} - F_i, \quad D_i^- = F_i - \frac{i-1}{n}, \quad \text{where } F_i = \Phi(z_{(i)}).$$

i	$x_{(i)}$	$z_{(i)}$	F_i	i/n	D_i^+	D_i^-
1	3	-0.705	0.2405	0.125	-0.116	0.241
2	4	-0.626	0.2655	0.250	-0.016	0.141
3	6	-0.470	0.3193	0.375	0.056	0.069
4	7	-0.392	0.3477	0.500	0.152	-0.027
5	8	-0.313	0.3771	0.625	0.248	-0.123
6	10	-0.157	0.4378	0.750	0.312	-0.187
7	16	+0.313	0.6229	0.875	0.252	-0.127
8	42	+2.349	0.9906	1.000	0.009	0.116

5. Interpretation

- $D = 0.312$ is driven by the gap at $x = 10$ ($i=6$), not by the extreme value $x = 42$. The right skew compresses the data's bulk into the lower range, causing the empirical CDF to outpace the fitted normal in the middle.

3. Test Statistic

$$D^+ = \max_i D_i^+ = 0.3122 \quad \text{at } i = 6, x = 10$$

$$D^- = \max_i D_i^- = 0.2405 \quad \text{at } i = 1, x = 3$$

$$D = \max(D^+, D^-) = \mathbf{0.3122}$$

At $i = 6$: the empirical CDF has risen to $6/8 = 0.750$ but the fitted normal CDF is only $\Phi(-0.157) = 0.438$ — a gap of 0.312. This signals the distribution's bulk lies lower than a normal would predict, consistent with right skew.

4. Results and Decision

Statistic	Value	CV (5%)	Decision
D	0.3122	0.288	Reject H_0

Lilliefors CVs (Abdi et. al. 2007) for $n = 8$:

α	Lilliefors CV	Standard KS CV	Decision
10%	0.265		Reject
5%	0.288	0.454	Reject
1%	0.333		FTR

$p \approx 0.021$. Rejects at 5% but not 1%.

Practitioner Insight: The KS statistic D measures the largest vertical gap between F_n and the fitted CDF, so only the single worst deviation drives the test. This makes it less powerful than Cramér–von Mises and Anderson–Darling, which use all observations. When μ and σ are estimated, the Lilliefors correction is required: standard KS critical values are too large and tend to under-reject, whereas Lilliefors uses smaller thresholds and is more appropriate. In skewed data, the maximum gap often occurs in the middle rather than at extremes due to shifts in the fitted distribution.

Common Pitfalls

- Using the K–S test for normality with estimated parameters.** Standard K–S critical values assume a fully specified distribution, so estimating μ and σ makes the test overly conservative; the Lilliefors correction must be used instead.
- Assuming the test is among the most powerful goodness-of-fit procedures.** Because the statistic depends only on the maximum deviation, it discards information about the rest of the distribution and is typically less powerful than tests such as Shapiro–Wilk or Anderson–Darling.
- Applying the test to discrete data or samples with many ties.** The K–S test assumes a continuous distribution, and ties distort the empirical distribution function, often making the test conservative and invalid under standard theory.
- Interpreting the statistic as a measure of overall fit.** The statistic D captures only the largest pointwise discrepancy, so it may miss widespread moderate deviations or be driven by a single outlier; graphical diagnostics are essential.
- Relying on asymptotic critical values in small samples.** The approximation $D_{\alpha,n} \approx c_\alpha/\sqrt{n}$ is inaccurate for small n , so exact critical values or software-based calculations should be used for reliable inference.

Jarque-Bera Test

Moment-Based Omnibus Test for Normality

NORM.6 | **Category:** Normality | **Data:** Continuous | **Null:** Chi-square

When to Use

- Test normality in large samples ($n > 50$, preferably $n > 100$).
- Quick computational check when only moments (skewness and kurtosis) are of concern.

Assumptions

- Observations are independent and identically distributed.
- Data have finite variance and fourth moment.
- Asymptotic approximation requires adequately large sample.

Test Statistic

$$JB = \frac{n}{6} \left[S^2 + \frac{(K-3)^2}{4} \right] \sim \chi^2_2 \text{ approximately under } H_0$$

where $S = m_3/m_2^{3/2}$ and $K = m_4/m_2^2$, with $m_k = (1/n) \sum (x_i - \bar{x})^k$.

Notation: S is sample skewness, K is sample kurtosis, normal has skewness = 0, kurtosis = 3.

Plain English: The Jarque–Bera test evaluates normality by measuring departures from the normal distribution’s defining shape characteristics—symmetry and tail thickness (kurtosis). It combines information from sample skewness and kurtosis to assess whether the observed distribution is consistent with what would be expected under normality.

⚠ Practitioner Warning

The Jarque–Bera test relies on asymptotic approximations and assesses normality using only skewness and kurtosis, making it insensitive to other distributional features such as multimodality or localized deviations. As a result, it can exhibit poor power in small samples and may fail to detect important departures from normality. In modern practice, Shapiro–Wilk or Anderson–Darling tests are often preferred for diagnostic purposes, while Jarque–Bera remains useful as a simple, moment-based summary check.

Hypotheses

- H_0 : Data distribution is Normal.
- H_1 : Data distribution is **not** Normal.

Decision Rule: Reject H_0 if $JB > \chi^2_{2,\alpha}$ or $p\text{-value} < \alpha$.
 $\chi^2_{2,0.01} = 9.210$, $\chi^2_{2,0.05} = 5.991$, $\chi^2_{2,0.10} = 4.605$.

📦 Software References [Python, R, SAS]

`scipy.stats.jarque_bera`

`tseries::jarque.bera.test`

PROC AUTOREG (NORMAL)

Additional Notes

The Jarque–Bera test is a moment-based normality test that decomposes departures from normality into contributions from skewness and excess kurtosis, making its diagnostic interpretation straightforward. The reference χ^2_2 distribution is asymptotic and may be inaccurate in small samples, motivating finite-sample corrections (e.g., Urzúa’s adjustment - given below) or simulation-based calibration. Because the test depends only on low-order moments, different non-normal distributions can yield similar Jarque–Bera statistics.

$$JB_U = \frac{n}{6} \left(\frac{S^2}{\text{Var}(S)} + \frac{(K-3)^2}{\text{Var}(K)} \right), \quad \text{Var}(S) = \frac{6(n-2)}{(n+1)(n+3)}, \quad \text{Var}(K) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Miscellanea: The Jarque-Bera test exemplifies how statistical tests evolve through cumulative contributions. While Jarque and Bera (1980) are credited with the specific test statistic—derived as a Lagrange multiplier test in the regression context—the fundamental idea of combining sample skewness and kurtosis to test normality was explored by D’Agostino (1970-1973) and formalized graphically by Bowman and Shenton (1975) in their omnibus test contours.

Reference(s): Jarque & Bera (1980, 1987); Urzúa (1996); Bowman & Shenton (1975)

Worked Example

Problem Statement Ten daily return percentages are tested for normality using the Jarque-Bera test, which combines sample skewness and excess kurtosis into a single moment-based statistic. Unlike the order-statistic tests (SW, SF, AD, CvM, KS), JB tests specific moment departures from normality rather than the overall shape of the empirical distribution. H_0 : data are normally distributed ($S = 0, K = 3$); H_1 : they are not.

1. Central Moments (divisor $n = 10$)

$\bar{x} = 13.0$; deviations $d_i = x_i - 13$:

x_i	d_i	d_i^2	d_i^3	d_i^4
3	-10	100	-1 000	10 000
5	-8	64	-512	4 096
6	-7	49	-343	2 401
7	-6	36	-216	1 296
9	-4	16	-64	256
10	-3	9	-27	81
12	-1	1	-1	1
15	+2	4	+8	16
21	+8	64	+512	4 096
42	+29	841	+24 389	707 281
Σ		1 184	22 746	729 524

$$m_2 = \frac{1184}{10} = 118.4, \quad m_3 = \frac{22746}{10} = 2274.6$$

$$m_4 = \frac{729524}{10} = 72952.4$$

2. Skewness and Kurtosis

$$S = \frac{m_3}{m_2^{3/2}} = 1.766, \quad K = \frac{m_4}{m_2^2} = 5.204$$

$$K - 3 = 2.204 \quad (\text{excess kurtosis})$$

3. JB Statistic Under H_0 : $JB \xrightarrow{d} \chi^2(2)$.

$$JB = \frac{n}{6} S^2 + \frac{n}{6} \frac{(K - 3)^2}{4} = 7.219$$

4. Results and Decision

Statistic	Value	CV (5%)	Decision
JB	7.219	5.991	Reject H_0

$p = 0.027$; CV at 1% = 9.210. Rejects at 5% but not 1%.

5. Contribution Breakdown

Component	Value	% of JB
Skewness	5.195	72%
Kurtosis	2.024	28%
Total JB	7.219	100%

Practitioner Insight: The Jarque-Bera test jointly assesses $S = 0$ (symmetry) and $K = 3$ (normal tails), with $JB \sim \chi^2(2)$ asymptotically from standardized skewness and kurtosis terms (the 1/4 weight reflects the larger variance of K). Here skewness dominates, driven by a single extreme value that inflates both m_3 and m_4 , producing positive skewness and excess kurtosis. While easy to compute and interpretable, JB is unreliable for small n (anti-conservative for $n < 50$); for larger samples it performs well, and for small samples a bootstrap p -value is preferable.

Common Pitfalls

- **Using the Jarque-Bera test with small samples.** The chi-square approximation is unreliable for small n , leading to low power and frequent non-rejection even for clearly non-normal data; alternative tests such as Shapiro-Wilk are preferred in small samples.
- **Assuming the test detects all forms of non-normality.** Because JB is based only on skewness and kurtosis, it can miss distributions that differ in shape but share the same third and fourth moments, so graphical diagnostics remain essential.
- **Over-interpreting the contribution of skewness and kurtosis components.** Although the statistic can be decomposed into skewness and kurtosis terms, these sample moments are correlated, so the decomposition provides only suggestive—not definitive—evidence about the source of deviation.
- **Ignoring sensitivity to outliers.** The reliance on higher-order moments makes the test highly sensitive to extreme values, where even a single outlier can dominate the statistic and drive rejection.

D'Agostino–Pearson K^2 Test

Omnibus Test Using Transformed Skewness and Kurtosis

NORM.7 | **Category:** Normality | **Data:** Continuous | **Null:** Chi-square

When to Use

- Test normality using refined skewness and kurtosis measures.
- Preferred over Jarque–Bera when better small-sample performance is needed.
- Suitable for moderate to large samples ($n \geq 20$).

Assumptions

- Observations are independent and identically distributed.
- Data have finite variance and fourth moment.
- Sample size should be at least $n \geq 20$ for adequate approximation.

Test Statistic

$$K^2 = Z_1^2 + Z_2^2$$

where transformations normalize skewness (b_1) and kurtosis (b_2) to approximate standard normal under H_0 . Under H_0 , $K^2 \sim \chi^2_2$ approximately. Specific transformations account for sample size and bias.

Notation: Z_1 is transformed skewness, Z_2 is transformed kurtosis, both approximately $N(0, 1)$

Plain English: D'Agostino–Pearson K^2 test improves upon Jarque–Bera by using more sophisticated transformations of sample skewness and kurtosis. These transformations better approximate normality under H_0 , particularly for moderate sample sizes, yielding more accurate p -values and better power than the simpler Jarque–Bera approach.

⚠ Practitioner Warning

The D'Agostino–Pearson K^2 test relies on asymptotic normal-score transformations of skewness and kurtosis and may exhibit reduced power in small samples. Because it captures departures from normality only through low-order moments, it can miss localized or multimodal deviations. In modern practice, Shapiro–Wilk or Anderson–Darling tests are often preferred for general diagnostic use, while K^2 remains useful as an interpretable, moment-based summary test.

Hypotheses

- H_0 : Data distribution is Normal.
- H_1 : Data distribution is **not** Normal.

Decision Rule: Reject H_0 if $K^2 > \chi^2_{2,\alpha}$ or if $p\text{-value} < \alpha$.
 $\chi^2_{2,0.01} = 9.210$; $\chi^2_{2,0.05} = 5.991$; $\chi^2_{2,0.10} = 4.605$.

📦 Software References [Python, R, SAS]

`scipy.stats.normaltest`

`moments::agostino.test`

Not Native (Manual)

Additional Notes

The D'Agostino–Pearson K^2 test applies normal-score transformations to skewness and kurtosis, improving finite-sample behavior relative to the Jarque–Bera test while retaining a χ^2_2 reference distribution. Because $K^2 = Z_1^2 + Z_2^2$, practitioners often examine Z_1 and Z_2 separately to identify whether departures from normality are driven primarily by asymmetry or tail behavior. The χ^2_2 reference distribution is asymptotic; for very small samples, simulation-based calibration may provide more accurate size control.

Miscellanea: The 1990 paper that established the K^2 test as a standard tool was co-authored by Ralph D'Agostino, Albert Belanger, and Ralph D'Agostino Jr.—making it a rare example of father-son collaboration in statistical methodology. The senior D'Agostino had developed the component tests decades earlier (1970–1973), while the younger D'Agostino, trained in biostatistics at Harvard, contributed modern computational validation and power comparisons.

Note: Normalization transformations are quite messy and take up space. They are provided in the Technical Notes of this chapter.

Reference(s): D'Agostino (1970, 1971); D'Agostino & Pearson (1973); D'Agostino et al. (1990)

Worked Example

Problem Statement Ten patient biomarker values are tested for normality using the D'Agostino–Pearson K^2 test. Unlike the Jarque–Bera test which uses raw skewness and excess kurtosis directly, K^2 first transforms each moment statistic into an approximately standard-normal score, then combines them. This makes the $\chi^2(2)$ approximation substantially more accurate for small n . H_0 : data are normally distributed; H_1 : they are not.

6, 7, 8, 9, 10, 11, 12, 16, 23, 38

1. Moments (divisor n)

$\bar{x} = 14.0$; deviations $d_i = x_i - 14$; $m_2 = \sum d_i^2/n = 86.4$

$$m_3 = \frac{1}{n} \sum d_i^3 = 1326.6, \quad m_4 = \frac{1}{n} \sum d_i^4 = 34712.4$$

$$b_1 = \frac{m_3}{m_2^{3/2}} = 1.652, \quad b_2 = \frac{m_4}{m_2^2} = 4.650$$

2. Z_1 : Skewness Transform (D'Agostino 1970)

$$Y = b_1 \sqrt{\frac{(n+1)(n+3)}{6(n-2)}} = 2.851$$

$$\beta_2(b_1) = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)} = 3.320$$

$$W^2 = -1 + \sqrt{2\beta_2 - 2} = 1.154, \quad \delta = \frac{1}{\sqrt{\ln \sqrt{W^2}}} = 3.734$$

$$\alpha = \sqrt{\frac{2}{W^2 - 1}} = 3.601, \quad Z_1 = \delta \cdot \operatorname{arcsinh}\left(\frac{Y}{\alpha}\right) = 2.712$$

3. Z_2 : Kurtosis Transform (Anscombe–Glynn 1983)

$$E[b_2] = \frac{3(n-1)}{n+1} = 2.455, \quad \sqrt{\operatorname{Var}[b_2]} = 0.755$$

$$x = \frac{b_2 - E[b_2]}{\sqrt{\operatorname{Var}[b_2]}} = 2.909$$

$$\beta_1(b_2) = 1.396, \quad A = 24.217$$

$$Z_2 = \sqrt{\frac{9A}{2}} \left[1 - \frac{2}{9A} - \left(\frac{1 - 2/A}{1 + x\sqrt{2/(A-4)}} \right)^{1/3} \right] = 2.175$$

4. Results

$$K^2 = Z_1^2 + Z_2^2 = 7.354 + 4.731 = 12.084$$

Stat	Value	CV (1%)	Decision
K^2	12.084	9.210	Reject H_0

$p = 0.002$. Rejects at 1%.

Practitioner Insight: The K^2 test improves on JB by applying normalizing transforms to skewness and kurtosis, yielding more accurate inference in small samples. Here JB ($p = 0.058$) fails to reject, while K^2 ($p = 0.002$) strongly rejects, because JB's $\chi^2(2)$ approximation is overly conservative when n is small. The D'Agostino and Anscombe–Glynn transforms map b_1 and b_2 to approximately $N(0, 1)$ variables, producing a better-calibrated test. As with JB, K^2 jointly detects skewness and kurtosis; here skewness dominates the departure from normality.

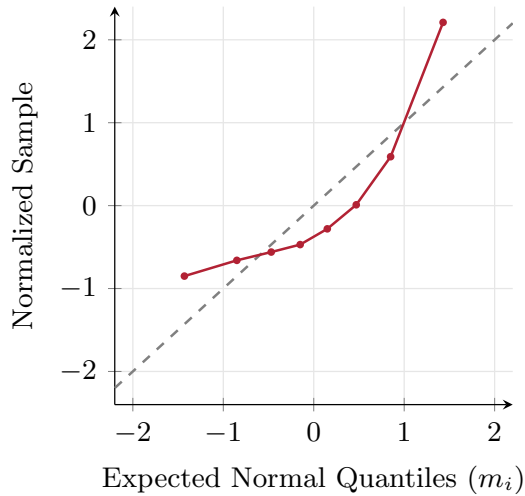
Common Pitfalls

- **Confusing D'Agostino's K^2 test with Jarque–Bera or chi-square goodness-of-fit tests.** D'Agostino's K^2 , Jarque–Bera, and chi-square goodness-of-fit tests are often conflated. K^2 uses transformed skewness and kurtosis statistics for improved approximation, whereas chi-square tests rely on binning and address a different hypothesis.
- **Using the test with very small samples.** The normalizing transformations rely on asymptotic approximations. For small n , performance deteriorates and more reliable inference typically requires moderate to large samples.
- **Assuming the test detects all forms of non-normality.** The test depends only on skewness and kurtosis. Distributions with correct moments but incorrect shape may not be detected, so graphical diagnostics remain essential.
- **Ignoring which component drives rejection.** A significant K^2 statistic can arise from skewness, kurtosis, or both. Inspecting the individual transformed components helps interpret the source of deviation.
- **Attempting manual computation of the transformations.** The transformations involve multiple steps and approximations. In practice, statistical software should be used to ensure correct and stable implementation.

I. NORM — Raw Data & Charts

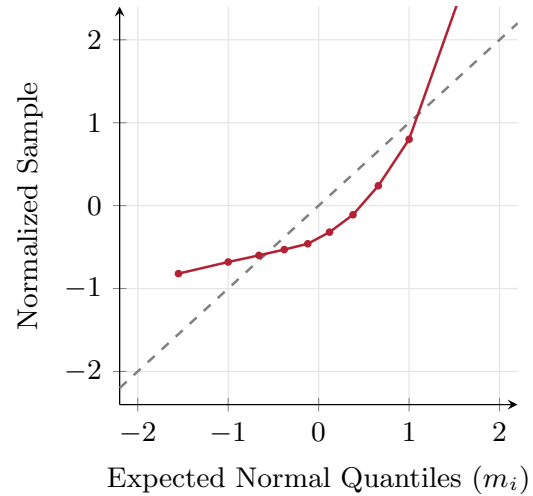
The raw data and corresponding Q–Q plots for all seven NORM tests are presented below. Each dataset is used in the worked example associated with its respective test.

Shapiro–Wilk Q–Q Plot



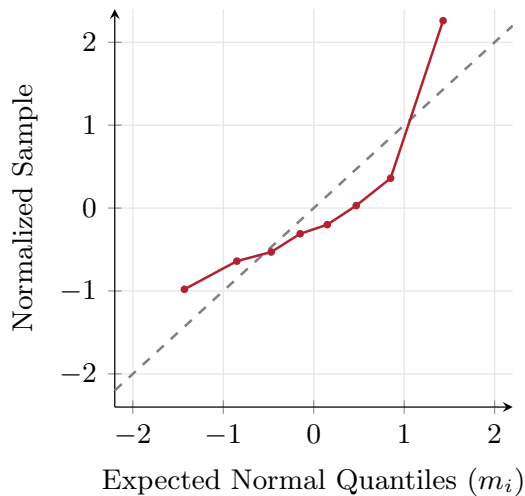
$X = \{3, 5, 6, 7, 9, 12, 18, 35\}$
 $W = 0.7936$, Reject H_0

Shapiro–Francia Q–Q Plot



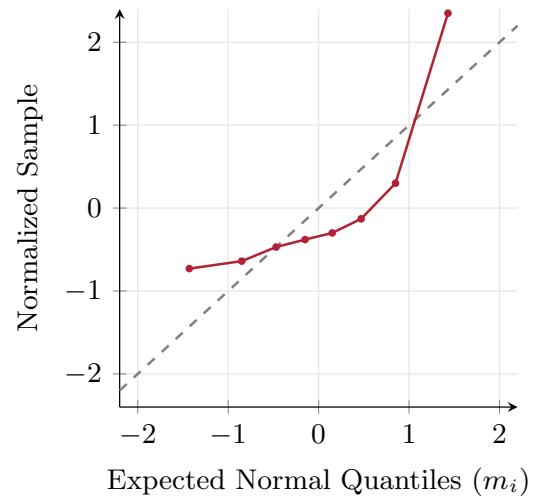
$X = \{5, 7, 8, 9, 10, 12, 15, 20, 28, 52\}$
 $W' = 0.7476$, Reject H_0

Anderson–Darling Q–Q Plot

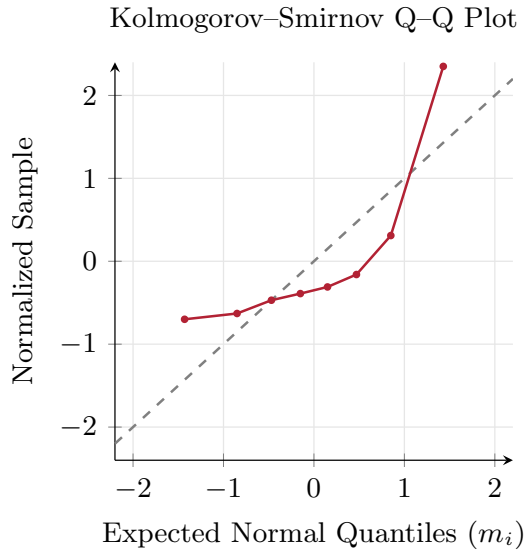


$X = \{10, 13, 14, 16, 17, 19, 22, 39\}$
 $A^{*2} = 0.776$, Reject H_0

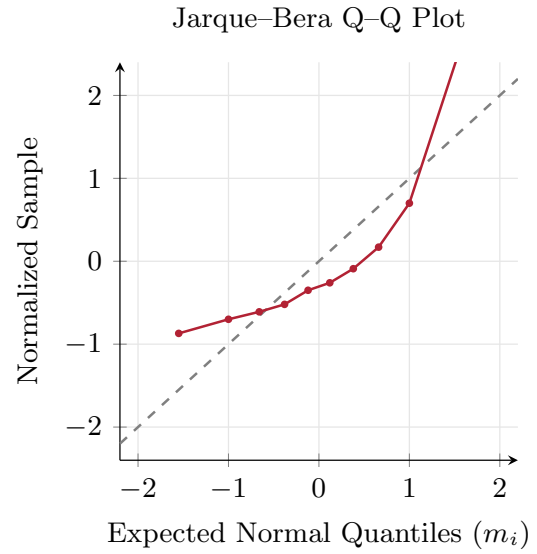
Cramér–von Mises Q–Q Plot



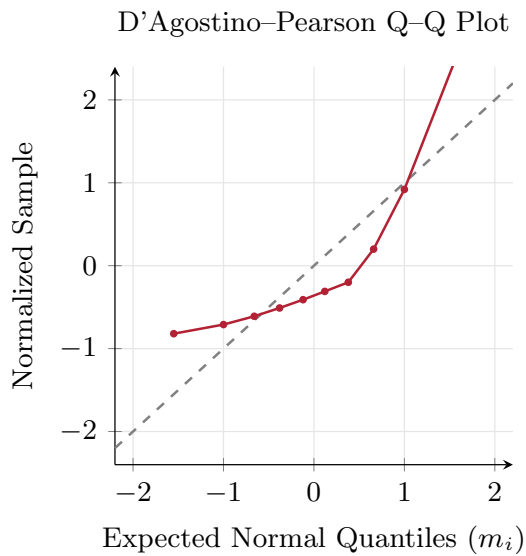
$X = \{4, 5, 7, 8, 9, 11, 16, 40\}$
 $W^{*2} = 0.1945$, Reject H_0



$X = \{3, 4, 6, 7, 8, 10, 16, 42\}$
 $D = 0.3122$, Reject H_0



$X = \{3, 5, 6, 7, 9, 10, 12, 15, 21, 42\}$
 $JB = 7.219$, Reject H_0



$X = \{6, 7, 8, 9, 10, 11, 12, 16, 23, 38\}$
 $K^2 = 12.084$, Reject H_0

I. NORM — Technical Notes

D'Agostino's K^2 Test Formulation

The D'Agostino's K^2 omnibus test assesses normality by combining transformations of sample skewness (S) and kurtosis (K) into a single χ^2 statistic.

Step 1: Compute Sample Moments

Given a sample of size n , compute the k -th central moments $m_k = \frac{1}{n} \sum (x_i - \bar{x})^k$:

$$\text{Skewness: } b_1 = \frac{m_3}{m_2^{3/2}}, \text{ Kurtosis: } b_2 = \frac{m_4}{m_2^2}$$

Step 2: Transform Skewness to $Z_1 \sim N(0, 1)$

$$Y = b_1 \sqrt{\frac{(n+1)(n+3)}{6(n-2)}}$$

$$\beta_2(b_1) = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$W^2 = -1 + \sqrt{2\beta_2(b_1) - 1}$$

$$\delta = \frac{1}{\sqrt{\ln W}}, \quad \alpha = \sqrt{\frac{2}{W^2 - 1}}$$

$$Z_1 = \delta \cdot \operatorname{arcsinh}\left(\frac{Y}{\alpha}\right), \text{ or}$$

$$Z_1 = \delta \cdot \ln\left(\frac{Y}{\alpha} + \sqrt{\left(\frac{Y}{\alpha}\right)^2 + 1}\right)$$

Step 3: Transform Kurtosis to $Z_2 \sim N(0, 1)$

$$E[b_2] = \frac{3(n-1)}{n+1}$$

$$\operatorname{Var}[b_2] = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

$$X = \frac{b_2 - E[b_2]}{\sqrt{\operatorname{Var}[b_2]}}$$

$$\sqrt{\beta_1(b_2)} = \frac{6(n^2 - 5n + 2)}{(n+7)(n+9)} \sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}} \left(\frac{2}{\sqrt{\beta_1(b_2)}} + \sqrt{1 + \frac{4}{\beta_1(b_2)}} \right)$$

$$Z_2 = \sqrt{\frac{9A}{2}} \left[1 - \frac{2}{9A} - \left(\frac{1 - 2/A}{1 + X\sqrt{2/(A-4)}} \right)^{1/3} \right]$$

Step 4: The Omnibus Statistic

$$K^2 = Z_1^2 + Z_2^2 \sim \chi^2(2)$$

PART II

Model Fit & Adequacy Assessment

Evaluating agreement between observed data and assumed models

GENERAL GOODNESS-OF-FIT

II. GFIT — A Conceptual Primer

1. Conceptual Foundation

What Is Goodness-of-Fit?

- Tests whether observed data are consistent with a **specified probability distribution** $F_0(x)$.
- Compares the **empirical distribution** of the sample against the **theoretical model**.
- Central question: *Is this distributional model plausible for these data?*
- Distinct from normality testing (NORM chapter): GFIT applies to *any* target distribution—exponential, Poisson, uniform, gamma, user-specified, etc.

The Null and Alternative Hypotheses

$$H_0 : F(x) = F_0(x) \quad \forall x \quad (\text{data follow } F_0)$$

- H_1 is typically **omnibus** (any departure): differences in location, scale, shape, tail weight, skewness, or multimodality.
- Unlike most tests in this compendium, GFIT tests are **non-directional**—they detect *that* the model fails, not *how* it fails.
- Tests such as Anderson–Darling remain omnibus but place greater emphasis on tail discrepancies than KS or CvM.

2. Distribution Comparison Framework

The Empirical CDF

The **empirical CDF** (ECDF) is the foundation of most GFIT tests:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(x_i \leq x).$$

By the Glivenko–Cantelli theorem, $\sup_x |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$, so F_n is a consistent estimator of the true CDF F . Different GFIT tests measure different aspects of the discrepancy between F_n and F_0 .

The Fundamental Principle

Every GFIT procedure measures a discrepancy between the empirical distribution F_n and the hypothesized distribution F_0 . The tests differ primarily in how that discrepancy is quantified: maximum distance (KS), integrated distance (CvM), tail-weighted distance (AD), spacing irregularity, or deviations in grouped frequencies (χ^2).

3. Hypothesis Structure

Simple vs. Composite Hypothesis

- **Simple H_0** : all parameters of F_0 are *fully specified* (e.g. $\text{Exp}(\lambda = 2)$). Standard critical values apply.

- **Composite H_0** : one or more parameters are *estimated from the data* (e.g. testing for exponentiality without specifying λ). Because parameters are estimated from the same data being tested, the discrepancy between F_n and F_0 is artificially reduced. Thus, ECDF is biased toward F_0 , critical values shift substantially, and using simple-hypothesis tables is **anti-conservative**.
- Corrections exist for common distributions: Lilliefors (normal, exponential), modified AD and CvM tables. When in doubt, use a **parametric bootstrap** to obtain valid critical values.

4. Taxonomy of Methods

Classes of GFIT Tests

EDF-Based (Continuous Data)

- **Kolmogorov–Smirnov (KS)**: maximum vertical distance between F_n and F_0 .
- **Anderson–Darling (AD)**: KS weighted by $[F_0(1 - F_0)]^{-1}$; emphasises tails.
- **Cramér–von Mises (CvM)**: integrated squared distance; balanced sensitivity.

Frequency-Based (Discrete or Binned)

- **Pearson χ^2** : compares observed vs. expected cell counts; requires adequate expected counts (≥ 5).
- **Likelihood ratio (G -test)**: $G^2 = 2 \sum O_i \log(O_i/E_i)$; additive across subtables; often preferred for sparse tables because of its likelihood-based interpretation. Covered in CTAB.

Specialized

- **Probability integral transform (PIT)**: if $X \sim F_0$, then $U = F_0(X) \sim \text{Uniform}(0, 1)$; reduces any GFIT problem to uniformity testing.
- **Spacing tests**: based on gaps $D_i = U_{(i)} - U_{(i-1)}$ between uniform order statistics; sensitive to clustering.

5. Core Mechanics

Core Test Statistics

$$\text{KS:} \quad D = \sup_x |F_n(x) - F_0(x)|$$

$$\text{CvM:} \quad W^2 = \int (F_n - F_0)^2 dF_0$$

$$\text{AD:} \quad A^2 = \int \frac{(F_n - F_0)^2}{F_0(1 - F_0)} dF_0$$

$$\chi^2: \quad X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$G\text{-test:} \quad G^2 = 2 \sum O_i \log(O_i/E_i) \quad (\text{covered in CTAB chapter})$$

A convenient computational form for EDF statistics uses order statistics $x_{(1)} \leq \dots \leq x_{(n)}$:

$$D = \max_i \left\{ \left| \frac{i}{n} - F_0(x_{(i)}) \right|, \left| F_0(x_{(i)}) - \frac{i-1}{n} \right| \right\}.$$

6. Visual Diagnostics

The P-P Plot

The **probability–probability (P–P) plot** plots $F_0(x_{(i)})$ against i/n (or $(i-0.5)/n$). Deviations from the 45° line reveal where F_0 over- or under-estimates the observed frequencies.

- P–P plots are sensitive to **center** of the distribution.
- Q–Q plots are sensitive to **tails**.
- Use both: Q–Q for tail fit, P–P for bulk fit.

7. Performance & Power

Sensitivity and Power Comparison

Test	Sensitive to	Best used when
KS	Center	Quick general check
CvM	Overall	Balanced alternative
AD	Tails	Tail behavior critical
χ^2	Any (binned)	Discrete/categorical F_0

AD often exhibits higher power than KS and CvM, particularly when departures occur in the tails; it is the **default recommendation** when tail fit matters. χ^2 is the only option for purely discrete F_0 .

Role of Sample Size

- **Small n (< 30):** low power; even substantial departures may go undetected. Visual diagnostics are essential.
- **Large n (> 200):** high power; trivial, practically irrelevant deviations are flagged as significant.
- **Key principle:** statistical significance $\neq$ practical significance. Ask whether the deviation matters for the downstream analysis, not merely whether it is detectable.

8. Chi-Square Specific Practicalities

Degrees of Freedom — χ^2 Test

- $df = k - 1$ (simple H_0 ; k cells).
- $df = k - 1 - m$ (composite H_0 ; m parameters estimated from grouped data).
- If parameters are estimated from *ungrouped* raw data (e.g. via MLE before binning), the χ^2 approximation is not exact; df is between $k - 1 - m$ and $k - 1$. Use simulation or bootstrap for reliable p -values.

Binning Strategy for χ^2

- **Expected count rule:** ensure all expected counts $E_i \geq 5$ (Cochran’s guideline). Merge adjacent bins if needed; for small samples, $E_i \geq 3$ may be tolerated with caution.
- **Equal-probability bins (recommended):** choose bin boundaries so that $E_i = n/k$. This improves power and avoids sensitivity to arbitrary cutpoints.
- **Choosing number of bins k :** use a moderate number of bins (typically $k \approx 5\text{--}15$), balancing resolution with the $E_i \geq 5$ constraint.
- **Heuristic rules (use with caution):** $k \approx \sqrt{n}$ (square-root rule) or $k \propto n^{2/5}$ (asymptotic MISE-based rules). These ignore distributional shape and are not preferred in practice.
- **Key principle:** binning is part of the test design. Poor binning reduces power and can distort conclusions.

9. Decision & Action Layer

What To Do After Rejection

- **Diagnose first:** use Q–Q and P–P plots to identify whether the deviation is in the tails, center, skewness, or kurtosis.
- **Try alternative families:** e.g. if exponential fails, try gamma or Weibull; if normal fails, try t or log-normal.
- **Transform:** log, $\sqrt{\cdot}$, Box–Cox to induce approximate conformity with a simpler target family.
- **Use robust or nonparametric methods** downstream that do not require a specific distributional form.
- **Re-examine data quality:** outliers, measurement error, and data contamination are frequent causes of rejection.

10. Practitioner Guardrails & Closure

Common Pitfalls

- Using standard KS critical values when parameters are estimated (anti-conservative).
- Poor or arbitrary binning in χ^2 (results depend on bin boundaries).
- Over-relying on p -values; ignoring effect size.
- Treating rejection as *model failure*—a minor, inconsequential departure may cause rejection at large n .
- Ignoring visual diagnostics (Q–Q, P–P, histogram overlay).
- Using χ^2 for continuous data when EDF tests have higher power.

Bottom Line

GFIT tests assess how well a model matches data—not whether the model is “true.” All models are approximations; the goal is *adequacy for purpose*, not perfection. Combine formal tests with visual diagnostics, assess whether detected deviations are practically consequential, and correct for parameter estimation whenever the null distribution is not fully pre-specified.

Anderson–Darling Test

Goodness-of-Fit for Specified Distributions

GFIT.1 | **Category:** Goodness-of-Fit | **Data:** Continuous | **Null:** Varies by distribution

When to Use

- Test fit to any fully specified continuous distribution.
- Emphasize tail deviations from the null distribution.
- Alternative to Kolmogorov–Smirnov with higher power for tail departures.
- Suitable for sample sizes $n \geq 5$; works well across wide range.

Assumptions

- Observations are independent and identically distributed.
- Data are continuous (ties reduce test accuracy).
- Null distribution $F_0(x)$ is **fully specified** (all parameters known).
- For distributions with estimated parameters, use modified critical values.

Test Statistic

$$A^2 = -n - \sum_{i=1}^n \frac{2i-1}{n} [\ln F_0(X_{(i)}) + \ln(1 - F_0(X_{(n+1-i)}))]$$

Tail-weighting via $[F_0(1 - F_0)]^{-1}$ makes test sensitive to extreme values.

Notation: $F_n(x)$ is empirical CDF, $F_0(x)$ is specified CDF, weighting emphasizes tails

Plain English: The Anderson–Darling test measures weighted discrepancy between observed and hypothesized distributions, giving extra weight to observations in the tails. This version can test fit to any continuous distribution—not just normality. The tail emphasis makes it particularly valuable for distributions where extreme values matter.

⚠ Practitioner Warning

When distribution parameters are estimated from data, standard critical values are invalid. Use distribution-specific adjustments or modified critical values. Each distribution family (exponential, Weibull, gamma, etc.) has unique adjustments derived through simulation. While the Anderson–Darling test is powerful, it is highly sensitive to large sample sizes: even minor, practically insignificant deviations can yield p -values below conventional thresholds. For this reason, it is best used in conjunction with Q–Q plots rather than relying on the p -value alone.

Hypotheses

- H_0 : Data follow the specified distribution $F_0(x)$.
- H_1 : Data do **not** follow the specified distribution $F_0(x)$.

Decision Rule: Reject H_0 if $A^2 > A^2_{\alpha,n}$ (critical value specific to F_0) or $p\text{-value} < \alpha$.

📦 Software References [Python, R, SAS]

`scipy.stats.anderson`

`goftest::ad.test`

PROC UNIVARIATE / NORMAL

Additional Notes

Critical values are distribution-specific because the null distribution of A^2 changes with F_0 . For common distributions (exponential, Weibull, logistic, extreme value), critical values have been extensively tabulated. The test extends naturally to censored data, making it valuable in survival analysis and reliability engineering. Finite sample corrections for a few different distributions are given below:

$$A^{2*} = A^2(1 + 0.75/n + 2.25/n^2)_{\text{Normal}} ; A^2(1 + 0.6/n)_{\text{Exponential}} ; A^2(1 + 0.2/\sqrt{n})_{\text{Weibull}} ; A^2(1 + 0.25/n)_{\text{Logistic}}.$$

Miscellanea: Although introduced in 1952, the Anderson–Darling test was not widely used for decades because its critical values depend on the specific null distribution (normal, exponential, etc.), unlike the Kolmogorov–Smirnov test with its universal tables. The test became practically accessible after Stephens (1974) tabulated and standardized these distribution-specific critical values, which is why many practitioners view the 1970s as its effective adoption period. The method was later extended to the k -sample setting (Scholz and Stephens, 1987), allowing comparison of multiple samples from an unknown distribution. Its creators, T.W. Anderson and D.A. Darling, were leading figures in 20th-century statistics, with Anderson’s work on multivariate analysis remaining foundational to this day.

Reference(s): Anderson & Darling (1952); Stephens (1974, 1986)

Worked Example

Problem Statement A service system is claimed to have exponentially distributed service times with *known* rate $\lambda = 0.5$ (mean = 2 min). Eight observed service times (minutes) are recorded. The Anderson–Darling test (1952) assesses fit to the *fully specified* $\text{Exp}(0.5)$ distribution: all parameters are given, none estimated from data.

$H_0: F(x) = 1 - e^{-0.5x}$; $H_1: F$ is not $\text{Exp}(0.5)$.

i	$x_{(i)}$	$F(x_{(i)})$	$1 - F(x_{(n+1-i)})$	$2i - 1$	$\log(F(x_{(i)}))$	$\log(1 - F(x_{(n+1-i)}))$
1	0.3	0.1393	0.0004	1	-1.971	-7.80
2	0.8	0.3297	0.0101	3	-1.110	-4.60
3	1.4	0.5034	0.0550	5	-0.686	-2.90
4	2.2	0.6671	0.1738	7	-0.405	-1.75
5	3.5	0.8262	0.3329	9	-0.191	-1.10
6	5.8	0.9450	0.4966	11	-0.057	-0.70
7	9.2	0.9899	0.6703	13	-0.010	-0.40
8	15.6	0.9996	0.8607	15	0.000	-0.15

1. AD Formula

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) [\ln F(x_{(i)}) + \ln(1-F(x_{(n+1-i)}))]$$

For $\text{Exp}(\lambda)$: $F(x) = 1 - e^{-\lambda x}$, so $1 - F(x) = e^{-\lambda x}$ and

$$\ln(1 - F(x)) = -\lambda x$$

The second log-term is therefore exact: $-0.5 x_{(n+1-i)}$.

2. AD Statistic and Decision

$$A^2 = -8 - (-10.930) = 2.930$$

Statistic	Value	CV (5%)	Decision
A^2	2.930	2.492	Reject H_0

$p \approx 0.031$; CV at 1% = 3.880 (FTR at 1%).

CVs: D'Agostino & Stephens (1986, pg. 105), fully specified.

Practitioner Insight: The general-fit Anderson–Darling test differs from the normality version in three key ways. First, all parameters must be *fully specified* in advance; estimating them (e.g. $\hat{\lambda} = 1/\bar{x}$) leads to a modified statistic with smaller values and different critical thresholds. Second, by the probability integral transform, $U_i = F(x_{(i)}) \sim \text{Uniform}(0, 1)$ under H_0 , so the test reduces to assessing non-uniformity of the transformed data, regardless of the hypothesized continuous distribution. Third, critical values are much larger in the fully specified case (e.g. 2.492 vs. 1.341 at 5%) because the model has no flexibility to adapt to the data. In this example, the large sample mean and extreme observation $x_{(8)} = 15.6$ (with $F(x) = 0.9996$) drive the result, illustrating the test's strong tail sensitivity relative to the Kolmogorov–Smirnov test.

Common Pitfalls

- **Wrong critical values.** Anderson–Darling critical values are distribution-specific; values for normal, exponential, or Weibull models are not interchangeable.
- **Unadjusted parameter estimation.** When parameters are estimated from data, standard critical values are invalid; use distribution-specific adjusted versions or software corrections.
- **Ignoring tail sensitivity.** Rejection typically reflects tail mismatches rather than central fit. Always inspect Q–Q plots with emphasis on extremes.
- **Discrete data misuse.** The test assumes continuity; applying it to discrete data leads to distorted inference. Use discrete goodness-of-fit methods instead.
- **Overgeneralized conclusions.** The test evaluates fit to a single specified distribution; non-rejection does not imply optimality among alternatives.

Cramér-von Mises Test

Balanced EDF Test for Any Specified Distribution

GFIT.2 | **Category:** Goodness-of-Fit | **Data:** Continuous | **Null:** Varies by distribution

When to Use

- Test fit to any fully specified continuous distribution with balanced sensitivity.
- Alternative to Anderson-Darling when uniform weighting is desired.
- Suitable for sample sizes $n \geq 5$; reliable across wide range.

Assumptions

- Observations are independent and identically distributed.
- Data are continuous (ties reduce validity).
- Null distribution $F_0(x)$ is **fully specified** (all parameters known).
- For distributions with estimated parameters, use modified critical values.

Test Statistic

$$W^2 = n \int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 dF_0(x) = \frac{1}{12n} + \sum_{i=1}^n \left[\frac{2i-1}{2n} - F_0(x_{(i)}) \right]^2$$

Notation: $F_n(x)$ is empirical CDF, $F_0(x)$ is specified CDF, equal weights across distribution

Plain English: The Cramér-von Mises test measures the integrated squared distance between observed and hypothesized distributions, treating all regions equally. Unlike Anderson-Darling which emphasizes tails, this test is balanced—equally sensitive to departures anywhere in the distribution. It extends naturally to any continuous distribution family.

⚠ Practitioner Warning

Critical values depend on the distribution being tested. The same W^2 value has different significance for exponential vs. normal vs. Weibull distributions. Always use appropriate critical values for the hypothesized F_0 . When parameters are estimated, further adjustments are required. Finite sample corrections (Stephens (1974)) are needed.

Hypotheses

- H_0 : Data follow the specified distribution $F_0(x)$.
- H_1 : Data do **not** follow the specified distribution $F_0(x)$.

Decision Rule: Reject H_0 if $W^2 > W_{\alpha,n}^2$ (critical value for distribution F_0) or $p\text{-value} < \alpha$. Consult distribution-specific tables or use software for accurate critical values.

📦 Software References [Python, R, SAS]

`scipy.stats.cramervonmises`

`gofest::cvm.test`

PROC UNIVARIATE / NORMAL

Additional Notes

The Cramér-von Mises test uses an L^2 (squared) distance, measuring integrated deviation between F_n and F with equal weighting across the distribution. Under a fully specified continuous F , the probability integral transform (PIT) yields a distribution-free test based on $\text{Uniform}(0, 1)$; however, when parameters are estimated (e.g. in normality testing), modified critical values are required. Stephens (1974) provided the practical critical values and approximations that enabled routine use of the test in such settings. For a fully specified F_0 , asymptotic critical values are 0.347 ($\alpha = 0.10$), 0.461 ($\alpha = 0.05$), 0.581 ($\alpha = 0.025$), and 0.743 ($\alpha = 0.01$); for finite samples, use $W^{*2} = (W^2 - 0.4/n + 0.6/n^2)(1.0 + 1.0/n)$ with these thresholds.

The Cramér-von Mises test is closely related to Watson's U^2 test for circular data, which modifies the statistic to remove dependence on the choice of origin, making it invariant to rotation—a key requirement for directional and periodic data.

Miscellanea: The Cramér-von Mises test has independent origins: Harald Cramér (1928) and Richard von Mises (early 1930s) arrived at essentially the same integrated squared-distance idea from different directions. Its practical development, however, is largely due to Nikolai Smirnov, who derived the sampling distributions and produced usable tables, establishing the test alongside Kolmogorov–Smirnov methods. In recognition of this, some refer to the broader class as the Cramér–von Mises–Smirnov family.

Reference(s): Cramér (1928); von Mises (1931); Anderson (1962); Stephens (1974)

Worked Example

Problem Statement A random-number generator is claimed to produce values uniformly distributed on $(0, 10)$. Six output values are recorded. $H_0: F(x) = x/10$ on $(0, 10)$; $H_1: F$ is not Uniform $(0, 10)$.

i	$x_{(i)}$	$F(x_{(i)})$	$(2i-1)/12$	d_i	d_i^2
1	0.5	0.05	0.083	0.033	0.001
2	1.0	0.10	0.250	0.150	0.023
3	1.5	0.15	0.417	0.267	0.071
4	2.0	0.20	0.583	0.383	0.147
5	2.5	0.25	0.750	0.500	0.250
6	9.5	0.95	0.917	-0.033	0.001

1. CvM Statistic

With $F(x) = x/10$ and $n = 6$:

$$\frac{2i-1}{2n} = \frac{2i-1}{12}, \quad F(x_{(i)}) = \frac{x_{(i)}}{10}, \quad d_i = \frac{2i-1}{2n} - F(x_{(i)})$$

$$W^2 = \frac{1}{12n} + \sum_{i=1}^n \left[\frac{2i-1}{2n} - F(x_{(i)}) \right]^2 = 0.507$$

2. Decision

Statistic	Value	CV (5%)	Decision
W^2	0.507	0.461	Reject H_0
W^{*2}	0.533	0.461	Reject H_0

$p \approx 0.036$; CV at 2.5% = 0.581 (FTR at 2.5%).

Critical values: Stephens (1974), fully specified.

Practitioner Insight: The CvM statistic $W^2 = 1/(12n) + n^{-1} \sum ((2i-1)/(2n) - U_{(i)})^2$, with $U_{(i)} = F(x_{(i)})$, measures the average squared deviation of probability-transformed data from their expected uniform positions. In this example, most U_i cluster in the lower tail, so the cumulative deviations dominate W^2 , while the single large observation contributes little. Unlike the Anderson–Darling test, CvM weights all deviations equally, yet both reach the same conclusion here because the lack of fit is spread across the bulk rather than concentrated in the tails. As with AD, fully specified models yield larger critical values since the model cannot adapt to the data.

Common Pitfalls

- **Wrong critical values.** Critical values depend on the distribution being tested; the null distribution of W^2 changes with F_0 . Always use values calibrated for the specific family.
- **No adjustment for estimated parameters.** Estimating parameters reduces W^2 , making standard critical values invalid. Use parameter-adjusted critical values (family-specific) when fitting θ from data.
- **Assuming universal power.** Because W^2 weights all deviations equally, it is balanced but not optimal. It may miss tail issues (AD) or moment deviations (SW). It is general-purpose, not most powerful.
- **Using with tied or discrete data.** The test assumes continuity; ties make it conservative. For inherently discrete data, use chi-square or discrete-specific GOF tests.
- **Interpreting non-rejection as confirmation.** Failing to reject only means compatibility with F_0 . Several distributions may fit; complement with AIC/BIC and graphical diagnostics.

Asymptotic Critical Values

For large n and fully specified F_0 : The limiting distribution can be expressed as

$$W^2 \xrightarrow{d} \int_0^1 [W(t)]^2 dt \stackrel{d}{=} \sum_{k=1}^{\infty} \frac{Z_k^2}{k^2 \pi^2}, \quad \text{where } W(t) \text{ is a Brownian bridge on } [0, 1] \text{ and } Z_k \sim N(0, 1) \text{ are independent.}$$

A Brownian bridge is a stochastic process obtained by conditioning Brownian motion to start and end at zero over a fixed interval (e.g. $[0, 1]$).

Kolmogorov-Smirnov Test (Two-Sample)

Distribution-Free Test for Comparing Two Samples

GFIT.3 | **Category:** Goodness-of-Fit | **Data:** Continuous | **Null:** Kolmogorov Distribution

When to Use

- Compare entire distributions of two populations.
- Detect location, spread, or shape differences.
- Distribution-free alternative to parametric tests.
- Continuous data from two independent samples.
- Compare distributions without assuming normality or any other specific forms.

Assumptions

- Both samples are independent and randomly selected.
- Observations within each sample are independent.
- Data are continuous (test is conservative with ties).
- No assumption about the underlying distribution.
- Suitable for any sample sizes; most powerful for moderate to large samples.

Test Statistic

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)| = \max \left\{ \max_i \left| \frac{i}{m} - G_n(x_i) \right|, \max_j \left| F_m(y_j) - \frac{j}{n} \right| \right\}$$

Notation: $F_m(x)$ and $G_n(x)$ are empirical CDFs from samples of sizes m and n .

Plain English: The two-sample Kolmogorov-Smirnov test compares two distributions by measuring the largest gap between their empirical cumulative distribution functions. If the samples come from the same distribution, their CDFs should be similar everywhere. A large gap anywhere indicates differing distributions, regardless of location, spread, or shape.

⚠ Practitioner Warning

K-S focuses on maximum deviation, which often occurs in the center of distributions. It can miss differences concentrated in the tails. For detecting tail differences, consider Anderson-Darling two-sample test or permutation tests. K-S is most powerful for location shifts and scale changes.

Hypotheses

- H_0 : $F(x) = G(x)$ Both samples come from same distn.
- H_1 : $F(x) \neq G(x)$ Samples come from different distn's.

Decision Rule: Reject H_0 if $D_{m,n} > D_{\alpha,m,n}$ or $p\text{-value} < \alpha$. For large samples, $D_{\alpha} \approx c_{\alpha} \sqrt{\frac{m+n}{mn}}$ where $c_{0.05} \approx 1.36$, $c_{0.01} \approx 1.63$.

📦 Software References [Python, R, SAS]

`scipy.stats.ks_2samp`

`ks.test`

PROC NPAR1WAY / EDF

Additional Notes

The test is distribution-free under the null hypothesis, with critical values depending only on the sample sizes m and n , not on the underlying distributions F or G . It is symmetric with respect to the two samples and invariant under strictly monotone transformations. The test is sensitive to any distributional difference, including shifts in location, scale, or shape, but summarizes all departures through a single global discrepancy, which can reduce power relative to parametric tests when their assumptions are satisfied.

The two-sample Kolmogorov-Smirnov test is distribution-free under H_0 , meaning its null distribution does not depend on the underlying continuous CDF; however, this property is lost under H_1 , making power analysis difficult. The statistic is invariant under all strictly increasing transformations, a rare robustness property that ensures identical results under re-scaling (e.g. log or Box-Cox transforms). Despite its generality, the test is most sensitive near the center of the distribution and relatively insensitive in the tails, motivating alternatives such as the Anderson-Darling and Cramér-von Mises tests.

Miscellanea: Historically, the method originated with Kolmogorov (1933) and was extended by Smirnov (1939), with early adoption limited by computational difficulty before modern methods made exact and approximate calculations routine.

Reference(s): Smirnov (1939); Hodges (1957).

Worked Example

Problem Statement Recovery times (days) are recorded for patients treated at two hospitals: Hospital A ($n_1 = 5$) and Hospital B ($n_2 = 6$). The two-sample KS test (Smirnov 1948) tests whether the two samples are drawn from the same continuous distribution, with no distributional form assumed. $H_0: F_1(x) = F_2(x)$ for all x ; $H_1: F_1(x) \neq F_2(x)$ for some x .

A: 1, 2, 3, 4, 8 B: 5, 6, 7, 9, 10, 11

1. Empirical CDFs at Each Combined Value

Pool and sort all $N = 11$ observations. At each value x :

$$\hat{F}_1(x) = \frac{\#\{\text{A obs} \leq x\}}{5}, \quad \hat{F}_2(x) = \frac{\#\{\text{B obs} \leq x\}}{6}$$

x	Src	$\hat{F}_1(x)$	$\hat{F}_2(x)$	$\hat{F}_1 - \hat{F}_2$	$ \hat{F}_1 - \hat{F}_2 $
1	A	1/5	0/6	+1/5	1/5
2	A	2/5	0/6	+2/5	2/5
3	A	3/5	0/6	+3/5	3/5
4	A	4/5	0/6	+4/5	4/5
5	B	4/5	1/6	+19/30	19/30
6	B	4/5	2/6	+7/15	7/15
7	B	4/5	3/6	+3/10	3/10
8	A	5/5	3/6	+1/2	1/2
9	B	5/5	4/6	+1/3	1/3
10	B	5/5	5/6	+1/6	1/6
11	B	5/5	6/6	0	0

2. Test Statistic

$$D = \sup_x |\hat{F}_1(x) - \hat{F}_2(x)| = \frac{4}{5} = 0.800$$

achieved at $x = 4$, where all four smallest combined observations belong to Hospital A while Hospital B has contributed nothing ($\hat{F}_2(4) = 0$).

3. Exact Critical Value and Decision

Under H_0 , all $\binom{11}{5} = 462$ ways of assigning the five A-labels to the eleven combined positions are equally likely. The exact tail probability is:

$$p = P(D \geq \frac{4}{5} \mid H_0) = \frac{22}{462} = \frac{1}{21} \approx 0.048$$

Statistic	Value	CV (5%)	Decision
D	0.800	0.800	Reject H_0

$p = 1/21 \approx 0.048$; D equals the exact 5% CV, so the test just rejects at the 5% level. Asymptotic CV: $1.358\sqrt{11/30} = 0.822$ (conservative for small n).

Practitioner Insight: The two-sample KS statistic $D = \sup_x |\hat{F}_1(x) - \hat{F}_2(x)|$ is the maximum vertical gap between the empirical CDFs, evaluated at the observed data points. Here, the gap of $4/5$ at $x = 4$ arises because all early observations from sample A precede those of sample B, after which the gap narrows as B accumulates. The test is distribution-free under H_0 , with the null distribution determined by permutations of sample labels, yielding the exact p -value. Unlike the one-sample case, the two-sample distribution depends on (n_1, n_2) , and asymptotic critical values can be conservative in small samples. Compared with AD and CvM, KS measures maximum pointwise deviation and is less sensitive to tail differences.

Common Pitfalls

- **Misusing K-S for location shifts.** When normality is plausible and only a mean shift is expected, a t -test is more powerful. Use K-S when shape or scale differences are possible or robustness is needed.
- **Weak tail sensitivity.** K-S often detects central rather than tail deviations. For tail-focused questions, prefer Anderson-Darling or tail-oriented permutation tests.
- **Misinterpreting rejection.** K-S is omnibus: rejection indicates some distributional difference, not specifically a mean shift. Use diagnostic plots to identify what differs.
- **Ignoring ties.** Ties make the test conservative. With rounded or discrete data, consider exact or permutation-based alternatives.
- **Small-sample approximations.** Asymptotic p -values are unreliable for small samples. Use exact or simulation-based methods when $\min(m, n)$ is small.

Lilliefors Test for Exponential Distribution

Modified K-S Test for Exponentiality with Estimated Rate

GFIT.4 | **Category:** Goodness-of-Fit | **Data:** Continuous | **Null:** Tables

When to Use

- Test whether data follow an exponential distribution with unknown rate parameter.
- Correct alternative to standard K-S when rate λ is estimated from data.
- Useful in reliability testing for constant failure rate assumption.

Assumptions

- Observations are independent and identically distributed.
- Data are continuous and positive (times, lifetimes, durations).
- Rate parameter λ is unknown and must be estimated.
- Suitable for sample sizes $n \geq 4$; better performance for $n \geq 20$.

Test Statistic

$$D = \sup_x |F_n(x) - \hat{F}_0(x)| = \max \left\{ \max_i \left| \frac{i}{n} - (1 - e^{-x_{(i)}/\bar{x}}) \right|, \max_i \left| (1 - e^{-x_{(i)}/\bar{x}}) - \frac{i-1}{n} \right| \right\}$$

Notation: $F_n(x)$ is empirical CDF, $\hat{F}_0(x)$ is exponential CDF with estimated rate $\hat{\lambda} = 1/\bar{x}$.

Plain English: The Lilliefors test for exponentiality checks whether data follow an exponential distribution when you don't know the rate parameter in advance. By estimating λ from the sample mean, the distribution is forced to fit the data better than if λ were fixed. This requires adjusted critical values that account for the parameter estimation, just like Lilliefors's correction for testing normality.

⚠ Practitioner Warning

The test has limited power for $n < 20$, often failing to detect departures from exponentiality. Power improves substantially with $n \geq 50$. For very small samples, consider graphical methods (exponential Q-Q plots) or combine with domain knowledge about the data-generating process. Critical values differ from standard K-S and are generally obtained via simulation for estimated λ .

Hypotheses

- H_0 : Data $\sim$ Exponential(λ), λ , unspecified
- H_1 : Data do **not** follow an exponential distribution.

Decision Rule: Reject H_0 if $D > D_{\alpha,n}$ (Lilliefors critical value for exponential) or if $p\text{-value} < \alpha$.

📦 Software References [Python, R, SAS]

statsmodels.stats.
diagnostics.lilliefors

nortest::lillie.test

PROC UNIVARIATE / NORMAL

Additional Notes

The test uses $\hat{\lambda} = 1/\bar{x}$ (method of moments estimator equals MLE for exponential). Critical values are about 15–25% smaller than standard K-S values for the same sample size. The exponential distribution is memoryless: $P(X > s + t | X > s) = P(X > t)$, making it appropriate for constant hazard rate processes.

Miscellanea: Hubert Lilliefors introduced the exponential version of his test in a brief 1969 note (3-page JASA article), extending his earlier normality work to reliability and life-testing where exponential waiting times are fundamental. Motivated by the practical flaw that the Kolmogorov–Smirnov test assumes known parameters, he showed how to adjust it when the rate λ is estimated, making the method usable in real data. Stephens (1974) later provided a simple normalization $T^* = (D - 0.2/n)(\sqrt{n} + 0.26 + 0.5/\sqrt{n})$, yielding approximate n -invariant critical values (about 1.094 at 5% and 1.298 at 1%), which made the test far easier to apply in practice. The method has since been used beyond engineering, including in seismology to assess whether earthquake inter-event times follow the exponential pattern implied by a Poisson process.

Reference(s): Lilliefors (1969); Dallal & Wilkinson (1986)

Worked Example

Problem Statement Eight machine inter-failure times (hours) are recorded. The Lilliefors test (1969) assesses whether the data follow *some* exponential distribution, estimating the rate λ by maximum likelihood ($\hat{\lambda} = 1/\bar{x}$) and then applying a KS-type statistic with critical values adjusted for the fact that λ is estimated rather than specified in advance. $H_0: F(x) = 1 - e^{-\lambda x}$ for some $\lambda > 0$; $H_1: F$ is not exponential.

1. MLE and Fitted CDF

$$\bar{x} = \frac{24.0}{8} = 3.0, \quad \hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{3}, \quad \hat{F}(x) = 1 - e^{-x/3}$$

2. KS Differences at Each Order Statistic

i	$x_{(i)}$	$\hat{F}(x_{(i)})$	$i/8$	$(i-1)/8$	D_i^+	D_i^-
1	0.1	0.0328	0.125	0.000	0.092	0.033
2	0.2	0.0645	0.250	0.125	0.186	-0.061
3	0.4	0.1248	0.375	0.250	0.250	-0.125
4	0.5	0.1535	0.500	0.375	0.347	-0.222
5	0.8	0.2341	0.625	0.500	0.391	-0.266
6	1.5	0.3935	0.750	0.625	0.357	-0.232
7	2.5	0.5654	0.875	0.750	0.310	-0.185
8	18.0	0.9975	1.000	0.875	0.003	0.123

3. Test Statistic and Stephens Modification

$$D^+ = \max_i D_i^+ = 0.391 \quad (\text{at } i = 5)$$

$$D^- = \max_i D_i^- = 0.123 \quad (\text{at } i = 8)$$

$$D = \max(D^+, D^-) = 0.391$$

Apply the Stephens (1974) modification to stabilize the critical values across sample sizes:

$$T^* = \left(D - \frac{0.2}{n}\right) \left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right) = 1.195$$

4. Decision

Statistic	Value	CV (5%)	Decision
T^*	1.195	1.094	Reject H_0

where $D_i^+ = i/n - \hat{F}(x_{(i)})$ and $D_i^- = \hat{F}(x_{(i)}) - (i-1)/n$.

$p < 0.05$; CV at 1% = 1.298 (FTR at 1%).

Critical values: Stephens (1974), exponential with estimated λ .

Practitioner Insight: Three features define the Lilliefors exponential test. First, the rate λ is *estimated* ($\hat{\lambda} = 1/\bar{x}$), so the fitted CDF adapts to the data and requires smaller critical values; the Stephens adjustment $T^* = (D - 0.2/n)(\sqrt{n} + 0.26 + 0.5/\sqrt{n})$ yields n -invariant asymptotic cutoffs (5%: 1.094; 1%: 1.298). Second, the largest deviation $D^+ = 0.391$ at $x_{(5)} = 0.8$ reflects that the sample accumulates mass much faster than the fitted exponential ($\hat{F}_n = 0.625$ vs. $\hat{F} = 0.234$), driven by the extreme value 18.0 inflating $\bar{x}$ and producing a distribution with heavier tails than exponential. Third, unlike the two-sample KS test, Lilliefors is a one-sample test against a parametric family (not another empirical CDF) and, while simpler than AD, is generally less sensitive to tail deviations since AD up-weights extremes.

Common Pitfalls

- **Wrong critical values with estimated λ .** Estimating $\lambda = 1/\bar{x}$ invalidates standard K-S critical values by shrinking D . Use Lilliefors critical values (or Stephens' adjustment) or the test becomes severely underpowered.
- **Invalid support.** The exponential distribution is defined only for $x > 0$. Zeros or negatives rule out exponentiality; consider shifted models or alternative families.
- **Low power for small n .** With $n < 20$, K-S-based tests have limited power. Always inspect exponential Q-Q plots and trust clear graphical departures over non-rejection.
- **Shifted or truncated data.** Standard tests assume support starting at zero. Known shifts must be removed; truncated samples require specialized methods.
- **Overlooking the hazard-rate assumption.** Exponentiality implies a constant hazard. Even if not rejected, assess whether this assumption is substantively reasonable for the application.

Kuiper's Test

Rotation-Invariant K-S Test for Circular and Linear Data

GFIT.5 | **Category:** Goodness-of-Fit | **Data:** Continuous | **Null:** Varies by distribution

When to Use

- Test fit to distributions on circular data.
- Alternative to Kolmogorov-Smirnov when rotation invariance is needed. Particularly effective for periodic data.
- Useful for linear data when sensitivity near distribution boundaries matters.

Assumptions

- Observations are independent and identically distributed.
- Data are continuous (or circular/periodic).
- For circular data, distribution is defined on a circle (angles, times on a cycle). Suitable for sample sizes $n \geq 5$.
- For linear data, null distribution $F_0(x)$ is fully specified.

Test Statistic

$$V = D^+ + D^- = \max_i \left(\frac{i}{n} - F_0(x_{(i)}) \right) + \max_i \left(F_0(x_{(i)}) - \frac{i-1}{n} \right)$$

Notation: D^+ is maximum upward deviation, D^- is maximum downward deviation of empirical from theoretical CDF.

Plain English: Kuiper's test is a modification of Kolmogorov-Smirnov that sums the maximum upward (D^+) and downward (D^-) deviations instead of taking their maximum. This makes the test rotation-invariant for circular data and equally sensitive to departures anywhere in the distribution, including at the boundaries—addressing a key weakness of K-S.

⚠ Practitioner Warning

Kuiper's test is less commonly implemented than K-S or Anderson-Darling. Specialized packages may be needed, particularly for circular data applications. For linear data, K-S or Cramér-von Mises may be more readily available with similar performance.

Hypotheses

- H_0 : Data follow the specified distribution F_0 (circular data)
- H_1 : Data do **not** follow the specified distribution F_0 .

Decision Rule: Reject H_0 if $V > V_{\alpha,n}$ or if $p\text{-value} < \alpha$. For large n , $V_{0.05,n} \approx 1.747/\sqrt{n}$; $V_{0.01,n} \approx 2.001/\sqrt{n}$.

📦 Software References [Python, R, SAS]

`astropy.stats.kuiper`

`circular::kuiper.test`

Not Native (Manual)

Additional Notes

Kuiper's test was introduced by Nicolaas Kuiper in 1960 while studying goodness-of-fit problems on the circle, particularly in astronomy. Its defining feature—rotation invariance—makes it ideal for circular data (e.g., angles, time-of-day), since the test statistic remains unchanged regardless of where the circle is “cut,” unlike the Kolmogorov-Smirnov test. This property made it especially valuable in fields like directional statistics and phase analysis.

The test gained prominence in astronomy through the analysis of periodic phenomena, such as pulsars and variable stars, where data naturally wrap around a cycle. In these applications, Kuiper's statistic is preferred over KS because it is equally sensitive across the entire domain, including near the boundaries—an area where KS tends to lose power.

Mathematically, Kuiper's statistic $V = D^+ + D^-$ can be viewed as a symmetrized version of the KS statistic, combining the largest positive and negative deviations. This makes it sensitive not only to shifts but also to changes in spread and shape. Despite its elegance, the test is less commonly taught than KS, partly because its critical values depend on more complex approximations, which historically limited its adoption before modern computing.

Miscellanea: Nicolaas Kuiper was a Dutch mathematician of remarkable range who later served as Director of the IHÉS in France. Although statisticians know him for Kuiper's test for circular uniformity, his deeper legacy lies in topology—most notably Kuiper's theorem showing that the unitary group of a separable infinite-dimensional Hilbert space is contractible.

Reference(s): Kuiper (1960); Stephens (1965)

Worked Example

Problem Statement Nine bird-migration departure directions (degrees clockwise from North) are recorded at a monitoring station. Kuiper's test (1960) assesses fit to the *fully specified* Uniform(0, 360) distribution—the null hypothesis that birds depart in all directions equally. Unlike the KS test, Kuiper's statistic $V = D^+ + D^-$ is invariant to the choice of origin on the circle, making it the correct analog for circular data. $H_0: F(\theta) = \theta/360$; H_1 : Directions are not uniformly distributed.

1. Probability Transform and Differences

Under H_0 : $u_i = \theta_{(i)}/360$. With $n = 9$, $i/n = i/9$ and $(i - 1)/n = (i - 1)/9$.

i	$\theta_{(i)}$	u_i	$i/9$	D_i^+	D_i^-
1	10	1/36	1/9	1/12	+1/36
2	20	1/18	2/9	1/6	-1/18
3	30	1/12	3/9	1/4	-5/36
4	40	1/9	4/9	1/3	-2/9
5	50	5/36	5/9	5/12	-11/36
6	70	7/36	6/9	17/36	-13/36
7	110	11/36	7/9	17/36	-13/36
8	130	13/36	8/9	19/36	-5/12
9	290	29/36	9/9	7/36	-1/12

where $D_i^+ = i/9 - u_i$ and $D_i^- = u_i - (i - 1)/9$.

2. Kuiper Statistic

$$D^+ = \max_i D_i^+ = 19/36 \quad (\text{at } i = 8, \theta = 130)$$

$$D^- = \max_i D_i^- = 1/36 \quad (\text{at } i = 1, \theta = 10)$$

All other $D_i^- < 0$: the empirical CDF runs ahead of the uniform CDF at every observation except $\theta = 10$.

$$V = D^+ + D^- = \frac{19}{36} + \frac{1}{36} \approx 0.5556$$

3. Stephens Modification

$$V^* = V \left(\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}} \right) \approx 1.797$$

4. Decision

Statistic	Value	CV (5%)	Decision
V^*	1.797	1.747	Reject H_0

$p < 0.05$; CV at 1% = 2.001 (FTR at 1%).

Critical values: Stephens (1970), fully specified.

Practitioner Insight (Condensed): Kuiper's statistic $V = D^+ + D^-$ combines the maximum upward and downward deviations, making it *rotation-invariant*: unlike KS, its value does not depend on where the circle is "cut." Here $D^+ = 19/36$ at 130° reflects strong clustering (8 of 9 observations in a narrow arc vs. 36% expected), while the lone outlier at 290° contributes little ($D^- = 1/36$). The choice $n = 9$ yields convenient exact fractions and $V = 5/9$. The Stephens adjustment $V^* = V(\sqrt{n} + 0.155 + 0.24/\sqrt{n})$ stabilizes critical values across n . Compared with the Rayleigh test (which is more powerful against unimodal concentration alternatives), Kuiper's test is more broadly sensitive to any departure from uniformity, including axial symmetry, multimodality, and gap patterns.

Common Pitfalls

- **Linear vs. circular misuse.** Kuiper is designed for circular data and rotation-invariant tests. For linear data, K-S (or A–D) is usually more powerful and better supported.
- **Incorrect handling of wrap-around.** Sorting circular data requires respecting angular continuity (e.g., 359° is near 1°). Software handles this; manual calculations must explicitly account for wrap-around.
- **Using with tied or discrete angles.** Kuiper assumes continuity; heavy rounding (e.g., nearest 10°) makes the test conservative. If discreteness is substantive, use discrete-specific circular tests.
- **Over-interpreting D^+ and D^- .** These components are not interpretable individually; the test statistic is the sum $V = D^+ + D^-$. Avoid directional over-interpretation.
- **Relying on asymptotic values for small n .** The approximation $V_\alpha \approx c_\alpha/\sqrt{n}$ is poor for $n \leq 20$. Use exact or simulation-based critical values for small samples.

Rao's Spacing Test

Circular Uniformity Test Based on Arc Lengths

GFIT.6 | **Category:** Goodness-of-Fit | **Data:** Continuous | **Null:** Simulated or Tables

When to Use

- Test uniformity of data on a circle (directions, angles).
- Detect clustering or concentration in circular data.
- Alternative to Rayleigh or Kuiper tests with different power characteristics.

Assumptions

- Observations are independent angular measurements. Data are measured on a circle (angles from 0° to 360°).
- No assumption about specific alternative distribution.
- Suitable for sample sizes $n \geq 4$; particularly effective for detecting clusters.

Test Statistic

$$U = \frac{1}{2} \sum_{i=1}^n |T_i - \lambda|, \quad \text{where } T_i = \theta_{(i+1)} - \theta_{(i)} \text{ with } \theta_{(n+1)} = \theta_{(1)} + 360.$$

Notation: T_i are arc lengths (spacings) between consecutive ordered angles, $\lambda = 360/n$ is expected uniform spacing.

Plain English: Rao's spacing test examines whether angles are uniformly distributed around the circle by checking if the gaps between consecutive points are roughly equal. If points cluster together, some gaps will be small and others large, yielding a large test statistic. The test is specifically designed for circular data and has excellent power for detecting clustering.

⚠ Practitioner Warning

Points must be ordered around the circle. Unlike linear data where ordering is natural, circular data can be ordered starting from any point. Software for circular statistics handles this automatically, but manual calculation requires careful attention to wrap-around from 360° back to 0°.

Hypotheses

- H_0 : Data are uniformly distributed on the circle.
- H_1 : Data are **not** uniformly distributed on the circle.

Decision Rule: Reject H_0 if $U > U_{crit}$ or if $p\text{-value} < \alpha$. Critical values from simulation or tabled values.

📦 Software References [Python, R, SAS]

Not Native (Manual)

`circular::rao.spacing.test`

Not Native (Manual)

Additional Notes

Rao's spacing test is based entirely on arc lengths between successive observations, making it inherently rotation-invariant and free of any parameter estimation; the null distribution depends only on n . The exact sampling distribution of U under uniformity is known but analytically intractable, so practical use relies on tabulated critical values or simulation. Rao (1976) provided tables for common levels ($p = 0.10, 0.05, 0.01$) for n up to 200, later extended by Russell & Levitin (1995) to $n = 1000$ and broader significance levels. Because the test is spacing-based rather than EDF-based, it is relatively robust to ties caused by rounding (e.g., integer-degree measurements), where KS-type methods may degrade. Modern implementations typically compute p -values via Monte Carlo simulation or refined approximations rather than relying solely on tables.

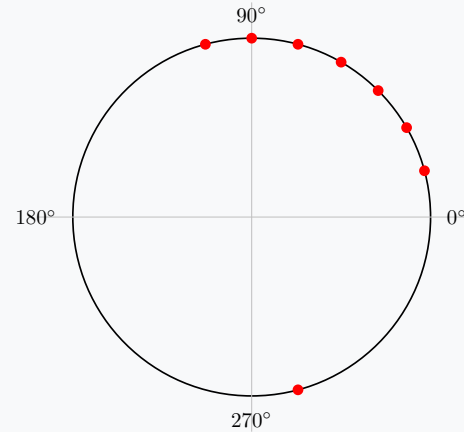
Miscellanea: Rao's spacing test emerged from the rich research environment of the Indian Statistical Institute (ISI) in the late 1960s, where J.S. Rao (his PhD supervisor was J. Sethuraman) developed the method in his dissertation under the broader influence of C.R. Rao's (Director of ISI) statistical legacy. The test was designed in part to overcome a key limitation of the popular Rayleigh test, which can miss multimodal patterns (e.g., data clustered in opposite directions). By focusing on spacings rather than vector sums, Rao's statistic detects bimodal and irregular clustering with ease. Geometrically, U measures the "departure from a perfect ring," quantifying how far observed points are from forming a regular polygon on the circle. Beyond its origins in directional data, the test has found modern applications in fields like chronobiology, where it is used to detect non-uniform activity patterns over a 24-hour cycle.

Reference(s): Rao, J. S. (1969, 1976); Russell & Levitin (1995)

Worked Example

Problem Statement Eight wind-direction readings (degrees clockwise from North) are recorded at a coastal station. Rao's spacing test (1976) assesses uniformity of circular data by examining whether the arcs between consecutive observations are consistent with a uniform distribution. Under H_0 each of the $n = 8$ spacings has expected length $\lambda = 360^\circ/8 = 45^\circ$. H_0 : Directions are uniformly distributed on $(0^\circ, 360^\circ)$; H_1 : Directions are not uniformly distributed.

i	$\theta_{(i)}$	$\theta_{(i+1)}$	$T_i = \theta_{(i+1)} - \theta_{(i)}$	$\frac{1}{2} T_i - \lambda $
1	15	30	15	15
2	30	45	15	15
3	45	60	15	15
4	60	75	15	15
5	75	90	15	15
6	90	105	15	15
7	105	285	180	67.5
8	285	15 + 360	90	22.5
Sum				180



Circular data for Rao's Spacing Test

1. Compute Spacings T_i

Arrange the $n = 8$ sorted angles around the circle. The i -th spacing is the arc from $\theta_{(i)}$ to $\theta_{(i+1)}$; the final spacing wraps around: $T_n = 360^\circ - \theta_{(n)} + \theta_{(1)}$.

2. Test Statistic

$$U = \frac{1}{2} \sum_{i=1}^n |T_i - \lambda| = 180^\circ$$

3. Decision

Statistic	Value	CV (5%)	Decision
U	180°	175.7°	Reject H_0

$p < 0.05$; CV at 1% $\approx 198^\circ$ (FTR at 1%).
Critical values from Monte Carlo (Rao 1976; 10^6 reps).

Practitioner Insight: Rao's statistic $U = \frac{1}{2} \sum |T_i - \lambda|$ measures the mean absolute deviation of arc lengths from $\lambda = 360^\circ/n$, with large U indicating clustering (many short spacings with compensating long gaps) and small U indicating regularity. Here, six tight 15° spacings and one large 180° gap reflect strong clustering. The identity $\sum |T_i - \lambda| = 360^\circ$ follows from $\sum T_i = 360^\circ$, which forces positive and negative deviations to balance. Unlike Kuiper's test, Rao's test penalizes each spacing, making it especially sensitive to clustering and also able to detect overly regular patterns. It is rotation-invariant and requires no parameter estimation, so its null distribution depends only on n .

Common Pitfalls

- **Wrap-around omission.** Failing to include the circular wrap-around spacing $T_n = (\theta_{(1)} + 360^\circ) - \theta_{(n)}$ treats angular data as linear and invalidates the test. Always verify that $\sum_i T_i = 360^\circ$ as a basic consistency check.
- **Applying the test to non-circular data.** Rao's spacing test is designed for angular or directional data. Mapping linear measurements into $[0, 360)$ does not make them circular; for linear data, use standard goodness-of-fit tests instead.
- **Over-interpreting the type of departure.** The test is omnibus: rejection indicates non-uniformity but does not identify whether the pattern is clustering, bimodality, or another structure. Inspect spacings and use circular plots (e.g., rose diagrams) to interpret the nature of the deviation.
- **Using chi-square approximations.** Rao's spacing statistic does not follow a chi-square distribution, even asymptotically. Always use tabulated or Monte-Carlo critical values specific to the sample size n .
- **Expecting power against smooth unimodal alternatives.** Rao's test is most powerful for clustering and multimodal patterns. For smooth unimodal concentration (e.g., von Mises alternatives), the Rayleigh test typically has higher power.

Greenwood Test for Uniformity

Spacing-Based Test for Uniformity on the Unit Interval

GFIT.7 | **Category:** Goodness-of-Fit Tests | **Data:** Continuous | **Null:** Simulated or Asymp. Normal

When to Use

- Test whether data follow a Uniform(0, 1) distribution.
- Assess p-value uniformity in simulation or meta-analysis.
- Evaluate randomness of pseudo-random number generators.
- Check uniform occurrence of rescaled event times on [0, 1].

Assumptions

- Independent, continuous observations on [0, 1].
- Under H_0 , data follow Uniform(0, 1).
- Sample size $n \geq 3$ (power improves with n).
- No ties (ties suggest discreteness or model violation).

Test Statistic

$$G = \sum_{i=1}^{n+1} (x_{(i)} - x_{(i-1)})^2, \quad x_{(0)} = 0, x_{(n+1)} = 1, \quad \left| \quad E[G] = \frac{2}{n+2}, \quad \text{Var}(G) = \frac{4n}{(n+2)^2(n+3)(n+4)} \right.$$

Notation: $x_{(1)} \leq \dots \leq x_{(n)}$ are order statistics, d_i are spacings between consecutive $x_{(i)}$, n is sample size (spacings= $n+1$)

Plain English: The Greenwood statistic assesses uniformity by examining how evenly observations partition the unit interval. Under uniform randomness, spacings are roughly $1/(n+1)$. Clustering creates uneven spacings and inflates G , while unusually regular spacing produces a small G . Either pattern signals deviation from uniformity.

Practitioner Warning

The test is two-sided: both large and small values of G can be significant, with critical regions in both tails. The null distribution depends on n and is typically obtained via tables, simulation, or asymptotic approximation. Data must be transformed to [0, 1] before application (p-values already satisfy this). Compared to EDF tests such as Kolmogorov–Smirnov and especially Anderson–Darling (which emphasizes tail discrepancies), Greenwood is less sensitive to tail deviations but may be more powerful for multimodal or spacing-driven departures.

Hypotheses

- H_0 : The observations are uniformly distributed on [0, 1].
 - H_1 : The observations are either clustered or overly regular.
- Decision Rule:** Reject H_0 if $G < G_{\alpha/2,n}$ or $G > G_{1-\alpha/2,n}$ where critical values are obtained using tables or simulation.

Software References [Python, R, SAS]

Not Native (Manual)

Not Native (Manual)

Not Native (Manual)

Additional Notes

The statistic arose from the idea that event times generated by a Poisson process, when properly scaled, should appear uniform on [0, 1]. Greenwood (1946) quantified deviation from this structure using spacing variability. Under H_0 , $E[G] = 2/(n+2)$ and $\text{Var}(G) = 4n/[(n+2)^2(n+3)(n+4)]$ (corrected by Moran). Moran (1947) showed that a standardized version of G is approximately normal for large n , though exact or simulated critical values are preferable in small samples.

Spacing-based tests differ geometrically from EDF tests: EDF methods measure vertical CDF discrepancies, while Greenwood measures horizontal spacing variability. This yields complementary power characteristics and links naturally to Rao's spacing test for circular data. In modern applications, Greenwood is frequently used to assess p-value uniformity in meta-analysis and large-scale testing.

Miscellanea: Major Greenwood, a pioneer of British medical statistics at the London School of Hygiene and Tropical Medicine, developed the spacing statistic while studying whether disease outbreaks clustered beyond chance expectation. He emphasized that visual inspection is unreliable for judging randomness, advocating formal statistical diagnostics instead. Originally computed by hand for small datasets, the statistic later found unexpected uses in cryptography and random number testing—demonstrating how a tool devised for epidemiology became a general instrument for detecting hidden structure in supposedly random sequences.

Reference(s): Greenwood (1946); Moran (1947); Pyke (1965); Stephens (1981); D'Agostino & Stephens (1986)

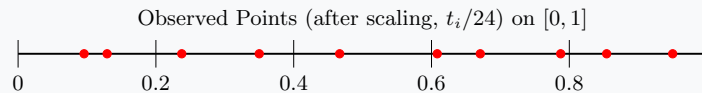
Worked Example

Problem Statement A quality control engineer is investigating whether defects occur uniformly along a continuous production line. The production line operates for 24 hours, and 10 defects were detected during this period. To assess whether these defects are uniformly distributed in time (which might suggest random occurrence) or show clustering (which might indicate systematic problems during certain periods), the engineer records the time of each defect occurrence.

Data: The sorted times (in hours from the start of the production period) at which the 10 defects occurred are:

$$U_i = (2.3, 3.1, 5.7, 8.4, 11.2, 14.6, 16.1, 18.9, 20.5, 22.8), \text{ for } i = 1, \dots, 10$$

Define $U_{(0)} = 0$, $U_{(11)} = 1$. The following plot shows the scaled points on the unit interval.



1. Greenwood statistic

$$G = \frac{1}{24^2} \sum_{i=1}^{n+1} (U_{(i)} - U_{(i-1)})^2 = \frac{1}{24^2} (2.3^2 + 0.8^2 + 2.6^2 + 2.7^2 + 2.8^2 + 3.4^2 + 1.5^2 + 2.8^2 + 1.6^2 + 2.3^2 + 1.2^2) = 0.1020$$

2. Critical Values and Decision (Monte Carlo, $n = 10$) Using 800,000 simulated samples under $H_0 : \text{Unif}(0, 1)$ (with endpoint spacings), the two-sided $\alpha = 0.05$ critical values are: $G_{0.025} = 0.1164$, $G_{0.975} = 0.2658$.

Observed $G_{\text{obs}} = 0.1020$, so $G_{\text{obs}} < G_{0.025}$ and we **reject** H_0 . Monte Carlo tail areas: $\hat{p}_L = \Pr(G^* \leq G_{\text{obs}}) \approx 0.00046$, $\hat{p}_U \approx 0.99954$, hence $\hat{p}_{2\text{-sided}} = 2 \min(\hat{p}_L, \hat{p}_U) \approx 0.00091$.

Practitioner Insight: Defect times are *not* consistent with uniform randomness over the 24-hour period; the unusually small G indicates *excessive regularity* (defects are more evenly spaced than a random uniform process would produce). This pattern is more consistent with a periodic/structured mechanism (e.g., scheduled cycles, batching, shift/maintenance effects, or reporting/inspection artifacts) than with random occurrence. The engineer should investigate time-structured drivers rather than clustering at specific hours.

Common Pitfalls

- **Failing to transform data to $[0, 1]$.** Greenwood tests uniformity. For other distributions, apply the probability integral transform first. Using raw data invalidates the test.
- **Ignoring dependence.** The test assumes independent observations. Serial correlation (e.g., time series or clustered events) distorts spacings and inflates Type I error.
- **Applying to discrete or tied data.** Ties violate the continuous uniform assumption and artificially shrink spacings, biasing G .
- **Using asymptotic normal approximation for small n .** The null distribution of G depends on n ; simulation or exact tables are preferable for moderate sample sizes.
- **Misinterpreting direction of departure.** Both unusually small and unusually large values of G indicate non-uniformity. Reporting only one tail misses half the diagnostic signal.
- **Confusing mixture alternatives with miscalibration.** In p-value analysis, non-uniformity may reflect genuine signals rather than flawed testing procedures.
- **Overlooking boundary effects.** Spacings include the intervals $(0, U_{(1)})$ and $(U_{(n)}, 1)$. Omitting endpoints changes the statistic.
- **Expecting strong tail sensitivity.** Greenwood is spacing-based and may miss alternatives concentrated only in extreme tails compared to Anderson–Darling.

Chi-Square Goodness-of-Fit Test

Pearson's Classic Test for Categorical Data

GFIT.8 | **Category:** Goodness-of-Fit | **Data:** Categorical/Discrete | **Null:** Chi-square

When to Use

- Test whether observed frequencies match expected frequencies from a specified distribution.
- Compare categorical data to a theoretical distribution or model.
- Appropriate for discrete distributions (binomial, Poisson, etc.) or grouped continuous data.

Assumptions

- Observations are independent. Total sample size is fixed.
- Data are counts in mutually exclusive categories.
- Expected frequency in each cell is at least 5 (preferably ≥ 5 in all cells).
- For estimated parameters, degrees of freedom must be adjusted.

Test Statistic

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}. \quad \text{Under } H_0, \chi^2 \sim \chi_{k-1-p}^2.$$

Notation: O_i are observed counts, E_i are expected counts under H_0 , k is number of bins, p is the number of parameters estimated

Plain English: The chi-square goodness-of-fit test compares observed counts to expected counts by measuring how much the two differ, scaled by what we'd expect from random sampling variation. Large discrepancies (relative to expected counts) suggest the data don't follow the hypothesized distribution. It's one of the oldest and most versatile tests in statistics.

⚠ Practitioner Warning

The chi-square approximation deteriorates when expected frequencies are small. Traditional rule: all $E_i \geq 5$. With smaller expected frequencies, the test becomes conservative (under-rejects). Combine sparse categories to meet this requirement, but avoid over-combining which loses information.

Hypotheses

- H_0 : Data follow the specified distribution
- H_1 : Data do **not** follow the specified distribution.

Decision Rule: Reject H_0 if $\chi^2 > \chi_{k-1-p, \alpha}^2$ or if $p\text{-value} < \alpha$.
df = $k - 1 - p$ where k is bins, p is estimated parameters.

📦 Software References [Python, R, SAS]

`scipy.stats.chisquare`

`chisq.test`

PROC FREQ / CHISQ

Additional Notes

The test is an asymptotic procedure for discrete data, with null distribution depending on the number of categories and any estimated parameters. The common guideline $E_j \geq 5$ is a rule of thumb rather than a strict requirement; smaller expected counts may be acceptable or can be handled by merging categories, though very small samples favor exact methods (e.g., Fisher's exact test). The test is asymptotically equivalent to the likelihood-ratio (G) test for multinomial models and can be viewed as a quadratic form of standardized residuals, linking it to modern diagnostics such as deviance residuals in generalized linear models. For continuous data, binning reduces power relative to EDF-based tests (KS, AD, CvM). For ordinal data; tests that account for ordering (e.g., Cochran–Armitage trend test) are more appropriate.

Miscellanea: Karl Pearson introduced the chi-square test in 1900 in a landmark paper often regarded as the birth of modern hypothesis testing. Originally motivated by problems in genetics, the test later played a central role in the famous Mendel–Fisher controversy: in 1936, R.A. Fisher showed that Mendel's pea-plant data were “too good to be true,” sparking debate that continues today. The test was also at the heart of a long-standing dispute between Pearson and Fisher, who corrected the degrees-of-freedom formula to account for estimated parameters—a refinement now standard. Despite its simplicity, the chi-square statistic has deep connections to the normal distribution: it arises as the sum of squared standard normal variables. Today, the test remains ubiquitous, from genetics (Hardy–Weinberg equilibrium) to machine learning feature selection, illustrating its enduring impact across scientific disciplines.

Reference(s): Pearson (1900); Cochran (1952)

Worked Example

Problem Statement A quality-control engineer records the number of surface defects on each of $n = 100$ circuit boards. The Poisson distribution is a standard model for defect counts, but the rate λ is unknown and must be estimated from the data. The chi-square goodness-of-fit test assesses whether the observed frequency distribution is consistent with *some* Poisson distribution. H_0 : counts follow $\text{Poisson}(\lambda)$ for some $\lambda > 0$; H_1 : counts do not follow any Poisson distribution.

j	$P(X=j)$	O_j	E_j	$O_j - E_j$	$(O - E)^2/E$
0	e^{-2}	22	13.53	+8.47	5.297
1	$2e^{-2}$	16	27.07	-11.07	4.525
2	$2e^{-2}$	26	27.07	-1.07	0.042
3	$\frac{4}{3}e^{-2}$	18	18.04	-0.04	0.000
4	$\frac{2}{3}e^{-2}$	12	9.02	+2.98	0.983
5+	$1 - F(4)$	6	5.27	+0.73	0.103
Sum	1	100	100	0	10.950

1. Estimate λ by MLE

For Poisson data, $\hat{\lambda} = \bar{x}$. Using the last grouped cell at $j = 5$ (approximation for $j \geq 5$):

$$\hat{\lambda} = \frac{\sum j O_j}{n} = 2.0$$

2. Expected Counts

$$E_j = n \cdot P(X=j; \hat{\lambda} = 2) = 100 \cdot \frac{2^j e^{-2}}{j!}$$

with $E_{5+} = 100 \cdot P(X \geq 5)$. All $E_j \geq 5$; total = 100.00 ✓

3. Chi-Square Statistic

$$X^2 = \sum_j \frac{(O_j - E_j)^2}{E_j} = 10.950$$

4. Degrees of Freedom and Decision

One parameter (λ) is estimated: $\text{df} = \text{bins} - 1 - p = 4$

Statistic	Value	CV (5%)	Decision
X^2	10.950	9.488	Reject H_0

$p = 0.027$; CV at 1% = 13.277 (FTR at 1%).

Practitioner Insight: The chi-square GoF test is unique in directly handling *discrete* distributions (e.g., Poisson, binomial), unlike EDF-based tests which require continuity. Its degrees of freedom $\text{df} = k - 1 - p$ adjust for parameter estimation, lowering critical values (e.g., from 11.07 to 9.49 here) and affecting rejection decisions. The $E_j \geq 5$ rule is a guideline—small cells can be merged. In this dataset, excess zeros and deficits at $j = 1$ suggest deviation from Poisson, consistent with a zero-inflated process.

Common Pitfalls

- **Small expected counts.** The chi-square approximation fails when any $E_i < 5$. Combine categories or use exact/simulation methods for small cells.
- **Incorrect degrees of freedom.** When parameters are estimated, use $\text{df} = k - 1 - p$. Ignoring the p parameter penalty gives invalid p -values.
- **Improper use with continuous data.** Chi-square requires categorical bins; binning loses information and results depend on cutpoints. Prefer EDF tests for continuous data.
- **Misreading omnibus rejection.** A significant result signals overall misfit, not which categories fail. Inspect individual contributions $(O_i - E_i)^2/E_i$ to locate problems.
- **Ignoring dependence.** The test assumes independent observations. For paired, clustered, or repeated data, use McNemar's test, Cochran's Q, or models accounting for dependence.

PART III

Location Differences & Central Tendencies

Testing contrasts in means, medians, and central location parameters

MEANS AND LOCATIONS

III. MEAN — A Conceptual Primer

1. Problem Framing

What Are We Testing?

- Whether the **central location** (mean, trimmed mean, or median) of one or more populations equals a specified value or differs across groups.
- Core question: *Are observed differences in sample means real or due to sampling chance?*

Hypotheses

One-sample: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ (or $>$, $<$ for directional tests).

Two-sample: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$.

k -sample: $H_0 : \mu_1 = \dots = \mu_k$ vs. H_1 : at least one μ_i differs.

2. Methods

The t -Family of Tests

All t -tests pivot on the same structure: standardize a mean difference by its estimated standard error and compare to a t -distribution.

One-sample t -test:

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

Paired t -test: define $d_i = x_{1i} - x_{2i}$; apply one-sample t to $\{d_i\}$:

$$T = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

Pairing removes between-subject variability; it **increases power** when observations are naturally matched.

Two-sample pooled t -test (equal variances):

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}, \quad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Welch's t -test (unequal variances; **default recommendation**):

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim t_\nu$$

where the Welch–Satterthwaite degrees of freedom are

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

Welch's test is **uniformly valid** whether or not variances are equal; prefer it unless equal variance is strongly justified by design or prior evidence.

One-Way ANOVA

Extends the two-sample comparison to $k \geq 2$ groups by decomposing total variability:

$$\underbrace{SS_{\text{Tot}}}_{= \sum_{ij} (y_{ij} - \bar{y})^2} = \underbrace{SS_B}_{k \text{ groups}} + \underbrace{SS_W}_{\text{within groups}}$$

$$F = \frac{MS_B}{MS_W} = \frac{SS_B/(k-1)}{SS_W/(N-k)} \sim F_{k-1, N-k} \quad \text{under } H_0.$$

Large F indicates between-group signal exceeds within-group noise. ANOVA tells you *that* at least one mean differs; it does not identify *which* groups differ.

Two-way ANOVA adds a second grouping factor (and possibly an interaction), partitioning SS_{Tot} further into main effects A , B , and $A \times B$.

3. Decision Layer

Choosing the Right Test

Design	Recommended test
1 sample vs. μ_0	One-sample t
2 matched/paired groups	Paired t
2 independent groups, equal var	Pooled t
2 independent groups, unequal var	Welch's t (default)
$k \geq 2$ groups, 1 factor	One-way ANOVA
k groups, 2 factors	Two-way ANOVA

4. Assumptions

Assumptions

Assumption	Consequence if violated
Independence	Biased p -values; invalid inference
Normality ($n < 30$)	Exact t/F distributions fail
Equal variances	Pooled t / ANOVA inflated Type I error

Robustness notes:

- t -tests are robust to mild non-normality, especially for $n \gtrsim 30$ (CLT); less so for heavy tails or outliers.
- ANOVA is robust to mild non-normality with balanced designs; *not* robust to unequal variances—use Welch ANOVA (Brown–Forsythe or Welch F) when variances differ.
- Independence violations are the most damaging; no correction from the t/F family will help—use mixed models (MIXD) or GEE for clustered/longitudinal data.

Statistical tests are foundational tools in modern analytical practice — yet finding the right one, understanding what it actually measures, and knowing when it breaks down requires hunting across textbooks, journal articles, and software documentation, often under time pressure.

The Compendium of Statistical Tests brings together over 250 tests in a single, consistent reference. Organized into 25 chapters spanning the full breadth of statistical practice — from classical parametric inference to Bayesian analysis, survival models, spatial statistics, multivariate analysis, and quality control — every test is presented in a standardized two-page format covering assumptions, test statistic, hypotheses, decision rule, worked example, common pitfalls, and software implementation in Python, R, and SAS.

Each chapter is preceded by a Conceptual Primer that builds intuition before the mathematics — because ideas come before the formulas.

For students, analysts, data scientists, statisticians, econometricians, clinical researchers, and engineers who engage with statistical models as part of their daily work, this compendium is the desk-side reference that has long been missing.

