

Sigma Statistics

Methodolgy

07th April 2026



Inhaltsverzeichnis

1. Normality Testing	5
1.1 Purpose and scope	5
1.2 Input data and preprocessing.....	5
1.3 Implemented tests and statistical definitions	5
1.3.1 Shapiro–Wilk test	5
1.3.2 Anderson–Darling test.....	6
1.3.3 Kolmogorov–Smirnov (Lilliefors-style normality test).....	6
1.3.4 D’Agostino–Pearson omnibus K^2 test.....	7
1.4 Hypotheses, significance level, and reported outputs	7
1.5 Assumptions and limitations	8
1.6 Warnings, edge cases, and safeguards.....	8
1.7 Numerical implementation and equivalence to R.....	8
1.8 References	9
2. Chi-square (χ^2) Test (2×2) and Fisher’s Exact Test	10
2.1 Purpose and scope	10
2.2 Input data, variable coding, and preprocessing	10
2.3 Contingency table notation	10
2.4 Pearson chi-square test with Yates’ continuity correction	11
2.5 Fisher’s exact test (two-sided)	11
2.6 Effect measures and confidence intervals	12
2.6.1 Odds ratio (OR)	12
2.6.2 Risk ratio (RR)	12
2.6.3 Risk difference (RD)	12
2.7 Low expected counts flag.....	13
2.8 Assumptions and limitations	13
2.9 Numerical implementation and equivalence to R.....	13
2.10 References	14
3. Unpaired (Independent Samples) t-Test	15
3.1 Purpose and scope	15
3.2 Input data and preprocessing.....	15
3.3 Descriptive statistics.....	15
3.4 Welch t-test (unequal variances)	15
3.5 Student t-test with pooled variance (equal variances assumed)	16
3.6 Confidence intervals for the mean difference	17
3.7 Effect sizes	17
3.8 Assumptions and limitations	17

3.9	Warnings, edge cases, and safeguards.....	18
3.10	Numerical implementation and equivalence to R.....	18
3.11	References.....	18
4.	Paired t-Test	19
4.1	Purpose and scope	19
4.2	Input data, pairing modes, and preprocessing.....	19
4.3	Numeric column acceptance (70% rule)	19
4.4	Definition of within-subject differences.....	20
4.5	Descriptive statistics.....	20
4.6	Paired t-test.....	20
4.7	Confidence interval for the mean difference	21
4.8	Effect sizes	21
4.9	Assumptions and limitations	21
4.10	Warnings, edge cases, and safeguards.....	22
4.11	Numerical implementation and equivalence to R.....	22
4.12	References	22
5.	Univariate Logistic Regression	23
5.1	Purpose and scope	23
5.2	Data structure, outcome definition, and predictors	23
5.3	Missing data handling and numeric checks.....	23
5.4	Logistic regression model.....	24
5.5	Estimation via IRLS (Newton–Raphson)	24
5.6	Standard errors, Wald test, odds ratio, and confidence interval.....	25
5.7	Warnings and quality safeguards	26
5.8	Assumptions and limitations	26
5.9	Numerical implementation and equivalence to R.....	27
5.10	References.....	27
6.	Multivariate Logistic Regression	28
6.1	Overview.....	28
6.2	Data Structure, Outcome Definition and Predictors	28
6.3	Missing Data, Numeric Checks and Listwise Deletion.....	28
6.4	Handling of Constant Predictors	29
6.5	Multivariate Logistic Regression Model	29
6.6	Estimation via IRLS (Newton–Raphson)	30
6.7	Standard Errors, Wald Tests and Confidence Intervals.....	30
6.8	Model-Level Diagnostics and Warnings	31
6.9	Predictor-Level Warning Codes.....	31
6.10	Assumptions and Limitations	32

6.11	Equivalence to R and numerical notes	32
7.	Propensity Score Matching (PSM).....	33
7.1	Purpose and scope	33
7.2	Notation and estimand.....	33
7.3	Inputs and variable types	33
7.4	Missing data handling (as implemented)	34
7.5	Propensity score model.....	34
7.6	Estimation (maximum likelihood via IRLS)	35
7.7	Distance measure for matching	35
7.8	Matching algorithm (nearest neighbor; 1:1; without replacement).....	36
7.9	Caliper matching.....	36
7.10	Overlap, common support, and unmatched units	37
7.11	Balance diagnostics (pre- and post-matching)	37
7.11.1	SMD definition used in Sigma.....	37
7.11.2	Pre- and post-matching samples	38
7.11.3	Additional descriptive balance metric.....	38
7.12	Interpreting balance	38
7.13	Outputs and exports.....	38
7.14	After matching: outcome analysis (principles).....	39
7.15	Assumptions and limitations	39
7.16	References.....	40
8.	Competing Risk Analysis — CIFs, Gray Test, and Fine & Gray Regression	41
8.1	Purpose and scope	41
8.2	Data structure and preprocessing	41
8.3	Event-time grid and basic risk set quantities	42
8.4	Overall survival for “any event”	42
8.5	Cumulative incidence functions (CIFs)	42
8.6	Variance and pointwise confidence intervals for CIFs	42
8.7	Gray’s test for group differences in CIFs (two groups; $\rho = 0$).....	43
8.8	Fine & Gray regression (subdistribution hazard model)	44
8.9	Outputs and exports.....	45
8.10	Assumptions and limitations	46
8.11	References.....	46

1. Normality Testing

1.1 Purpose and scope

The Sigma normality testing module assesses whether a univariate numeric sample is compatible with a Gaussian (normal) distribution. Normality testing is typically used to support methodological choices (e.g., whether a parametric model is plausible) and should be interpreted together with graphical diagnostics and the study context.

(Important: “Failing to reject normality” (e.g., $p \geq 0.05$) does not prove that the data are normal; it indicates that the observed deviations are not strong enough to reject normality at the chosen α . Conversely, with large sample sizes even small and clinically irrelevant deviations may yield $p < 0.05$.)

1.2 Input data and preprocessing

The module accepts one or more numeric variables, which are analyzed independently. Missing values and non-numeric entries are excluded on a per-variable basis (complete-case evaluation per variable).

All tests are computed on the resulting sample $x = (x_{1}, \dots, x_{n})$ of size n . Where required, the sample mean \bar{x} and sample standard deviation are estimated from the data. If all observations are identical or the variance is numerically zero, the respective test is not reported.

1.3 Implemented tests and statistical definitions

1.3.1 Shapiro–Wilk test

The Shapiro–Wilk test evaluates normality by measuring how closely the ordered sample resembles the expected order statistics of a normal distribution. Let $x_{(1)} \leq \dots \leq x_{(n)}$ denote the ordered sample and \bar{x} the sample mean. The test statistic is:

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where the coefficients are derived from expected values of normal order statistics and their covariance structure. The p-value is obtained using Royston-type approximations based on transformations of $1 - W$, following the standard Shapiro–Wilk computational approach.

(Sample size requirement: Sigma computes the Shapiro–Wilk test only for $3 \leq n \leq 5000$.)

1.3.2 Anderson–Darling test

The Anderson–Darling (AD) test is an EDF (empirical distribution function) goodness-of-fit test with relatively high sensitivity in the distribution tails. Let F denote the fitted normal CDF (using sample mean and standard deviation) and $x_{(i)}$ the ordered observations. The base AD statistic is:

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

Sigma reports the finite-sample adjusted statistic

$$A^{2*} = A^2 \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right),$$

and computes the p-value using Stephens-type piecewise approximations for the normality case.

(Sample size requirements: the AD statistic can be computed for $n \geq 3$. P-value reporting is based on the standard approximation and is provided for $n \geq 8$.)

1.3.3 Kolmogorov–Smirnov (Lilliefors-style normality test)

Sigma implements the one-sample Kolmogorov–Smirnov normality statistic against a fitted normal distribution, i.e. with mean and standard deviation estimated from the sample (composite null hypothesis). Let F be the fitted normal CDF and F_n the empirical CDF. The test statistic is:

$$D = \sup_x |F_n(x) - F(x)|.$$

For p-value computation, Sigma uses a Lilliefors-/Stephens-style approximation based on the transformed statistic

$$Z = D(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}),$$

followed by interpolation between standard tabulated reference points. Accordingly, the reported p-values are approximate p-values for the normality setting with estimated parameters, not the classical KS p-values for a fully specified distribution.

(Sample size requirement: this test is reported for $n \geq 5$.)

1.3.4 D'Agostino–Pearson omnibus K^2 test

The D'Agostino–Pearson omnibus test combines skewness and kurtosis into a single normality statistic. Sigma first computes skewness and excess kurtosis from sample central moments, transforms both components to approximately standard normal scores, and then forms:

$$K^2 = Z_{\text{skew}}^2 + Z_{\text{kurt}}^2.$$

Under the null hypothesis of normality, K^2 is evaluated against a χ^2 distribution with 2 degrees of freedom.

(Sample size requirement: due to the moment-based approximations, the test is computed for $n \geq 8$.)

1.4 Hypotheses, significance level, and reported outputs

For each selected test, the null hypothesis is H_0 : the data are drawn from a normal distribution (with parameters estimated from the sample where applicable). The alternative H_1 is that the distribution deviates from normality.

Sigma reports for each selected test:

- sample size n used by the test,
- the test statistic (e.g., W , A^{2*} , D , K^2),
- the corresponding p-value, if available for the implemented approximation regime.

(Default interpretation uses $\alpha = 0.05$ unless specified otherwise: $p < \alpha$ indicates evidence against normality, while $p \geq \alpha$ indicates that the sample is compatible with normality at that α level.)

1.5 Assumptions and limitations

- **Independence:** Observations are assumed independent.
- **Sensitivity to n:** With large n , normality tests may reject for small, practically irrelevant deviations; with small n , power can be limited.
- **Discrete / tied data:** Strong discreteness or many ties may affect EDF-based tests and moment-based approximations.
- **Degenerate data:** Constant or near-constant data can make one or more tests undefined.
- **Normality is not model adequacy:** Passing a normality test does not guarantee that a parametric model is appropriate; conversely, rejecting normality does not automatically invalidate parametric methods.

1.6 Warnings, edge cases, and safeguards

Sigma only reports a test if its minimum sample size requirement is met and the computation is numerically meaningful. Typical situations where results may be omitted or flagged include:

- n below the test-specific minimum,
- Shapiro–Wilk sample sizes above 5000,
- constant or near-constant data (variance approximately zero),
- invalid numeric input after preprocessing (e.g., all values missing/non-numeric).

1.7 Numerical implementation and equivalence to R

Computations are performed in JavaScript using double-precision floating point arithmetic (IEEE-754). Several p -values are based on established approximations rather than exact finite-sample distributions. Therefore, small differences between software packages may occur because of different approximation regimes, interpolation schemes, and rounding.

Conceptually, the procedures correspond to widely used reference implementations:

- `stats::shapiro.test` for Shapiro–Wilk,

- `nortest::ad.test`-type Anderson–Darling normality testing,
- `nortest::lillie.test`-type Lilliefors-corrected KS normality testing.

(In `Sigma`, the normality-module “Kolmogorov–Smirnov” result refers to the fitted-normal / Lilliefors-style variant, not to the classical KS test with completely fixed distribution parameters.)

1.8 References

- Royston, J. P. "W test for normality: algorithm AS 181." *Applied Statistics* 31.2 (1982): 176-180.
- Royston, P. "Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality." *Applied Statistics* 44.4 (1995): 547-551.
- Stephens, Michael A. "EDF statistics for goodness of fit and some comparisons." *Journal of the American Statistical Association* 69.347 (1974): 730-737.
- Lilliefors, Hubert W. "On the Kolmogorov-Smirnov test for normality with mean and variance unknown." *Journal of the American Statistical Association* 62.318 (1967): 399-402.
- R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Thode, Henry C. *Testing for normality*. CRC Press, 2002.

2. Chi-square (χ^2) Test (2×2) and Fisher's Exact Test

2.1 Purpose and scope

This module analyzes associations between a binary group variable (e.g., Control vs Treatment) and one or more binary outcome variables (parameters) using a 2×2 contingency table. For each outcome, Sigma reports both an approximate large-sample test (Pearson χ^2 with Yates' continuity correction) and an exact test (Fisher's exact test), together with common effect measures (odds ratio, risk ratio, risk difference).

2.2 Input data, variable coding, and preprocessing

The analysis requires:

- a group column with exactly two observed (non-missing) levels, and
- one or more parameter columns, each with exactly two observed (non-missing) levels.

Rows with missing values in the group column or the current parameter column are excluded for that parameter (complete-case evaluation per parameter). If the group column does not have exactly two levels, or a parameter column is not binary, the corresponding analysis is not performed.

The user selects which group level is labeled Control and which is labeled Treatment. For each parameter, Sigma forms a 2×2 table by mapping the two observed parameter levels to a reference level and an event level. Effect measures (OR/RR/RD) are reported for "Treatment vs Control" and "Event vs Reference".

(Note: The direction (sign) and magnitude of OR/RR/RD depend on which category is treated as "Event" versus "Reference". In Sigma, the event/reference assignment follows the two observed parameter levels as encountered in the data. If you need a specific event definition (e.g., "Death = 1"), ensure the category coding matches your intended direction.)

2.3 Contingency table notation

For each binary parameter, Sigma constructs the 2×2 table:

	Event	Reference	Total
Treatment	a	b	$a + b$
Control	c	d	$c + d$
Total	$a + c$	$b + d$	N

with total sample size $N = a + b + c + d$.

2.4 Pearson chi-square test with Yates' continuity correction

Under the null hypothesis of independence (no association between group and outcome), the expected counts are:

$$E_{ij} = \frac{(\text{row total})_i \cdot (\text{column total})_j}{N}$$

Sigma computes the Pearson chi-square statistic for a 2x2 table using Yates' continuity correction:

$$\chi_Y^2 = \sum_{\text{cells}} \frac{(\max(0, |O - E| - 0.5))^2}{E}$$

where O and E denote observed and expected counts for each cell. Degrees of freedom for a 2x2 table are $df = 1$.

The two-sided p-value is obtained from the chi-square distribution with 1 degree of freedom. For $df = 1$, Sigma uses the equivalent normal relationship:

$$p = 2(1 - \Phi(\sqrt{\chi_Y^2})),$$

where Φ is the standard normal CDF.

2.5 Fisher's exact test (two-sided)

Fisher's exact test evaluates the null hypothesis of independence conditional on the fixed row and column margins. Sigma computes the probability of the observed table under the hypergeometric model and reports a two-sided p-value by summing probabilities of all tables with the same margins whose probabilities are less than or equal to that of the observed table (standard two-sided definition for Fisher's exact test).

(In small samples or when expected counts are low, Fisher’s exact test is often preferred because it does not rely on large-sample chi-square approximations.)

2.6 Effect measures and confidence intervals

2.6.1 Odds ratio (OR)

The odds ratio compares the odds of the event in Treatment versus Control:

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

If any cell count is zero, Sigma applies the Haldane–Anscombe correction by adding 0.5 to all four cells before computing OR and its confidence interval.

A 95% confidence interval is computed using the Wald method on the log scale:

$$SE(\log OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}, CI_{95\%} = \exp(\log OR \pm 1.96 \cdot SE(\log OR)),$$

where the corrected cell counts are used if the Haldane–Anscombe adjustment was applied.

2.6.2 Risk ratio (RR)

The risk ratio compares event probabilities:

$$risk_T = \frac{a}{a+b}, risk_C = \frac{c}{c+d}, RR = \frac{risk_T}{risk_C}.$$

A 95% confidence interval is computed on the log scale (Wald method) when both event counts are nonzero:

$$SE(\log RR) = \sqrt{\left(\frac{1}{a} - \frac{1}{a+b}\right) + \left(\frac{1}{c} - \frac{1}{c+d}\right)}, CI_{95\%} = \exp(\log RR \pm 1.96 \cdot SE(\log RR)).$$

If $a = 0$ or $c = 0$, $\log RR$ is undefined; in this case Sigma may omit the RR CI.

2.6.3 Risk difference (RD)

The risk difference is:

$$RD = risk_T - risk_C.$$

A 95% Wald confidence interval is computed as:

$$SE(RD) = \sqrt{\frac{risk_T(1 - risk_T)}{a + b} + \frac{risk_C(1 - risk_C)}{c + d}}, CI_{95\%} = RD \pm 1.96 \cdot SE(RD).$$

2.7 Low expected counts flag

Sigma flags low expected counts if at least one expected cell count is < 5 . This is a common rule-of-thumb indicating that large-sample chi-square approximations may be unreliable, and that Fisher's exact test (or alternative methods) should be emphasized.

2.8 Assumptions and limitations

- Independence: observations are assumed independent; the test is not designed for paired or clustered data.
- 2x2 only: Sigma's χ^2 module evaluates only binary group vs. binary outcome (no general $r \times c$ tables).
- Direction depends on coding: OR/RR/RD depend on which outcome category is treated as "Event". Ensure coding matches the intended interpretation.
- Multiple testing: when many parameters are tested, Sigma does not automatically apply multiplicity corrections (e.g., Bonferroni or FDR); interpretation should consider the number of tests.

2.9 Numerical implementation and equivalence to R

Calculations are performed in JavaScript using double-precision floating point arithmetic (IEEE-754).

Conceptually, the procedures correspond to common reference implementations in R:

- `stats::chisq.test` (Pearson χ^2 ; note that Sigma applies Yates' correction for 2x2 tables),
- `stats::fisher.test` (two-sided Fisher's exact test).

(Small numerical differences between software packages may occur due to rounding and implementation details, but the underlying statistical definitions are standard.)

2.10 References

- Pearson, K. (1900). On the criterion that a given system of deviations from the probable... Philosophical Magazine 50(302): 157–175.
- Yates, F. (1934). Contingency tables involving small numbers and the chi-square test. Journal of the Royal Statistical Society (Suppl.) 1: 217–235.
- Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85(1): 87–94.
- Woolf, B. (1955). On estimating the relation between blood group and disease. Annals of Human Genetics 19: 251–253.
- Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. Annals of Human Genetics 20: 309–311.
- Anscombe, F. J. (1956). On estimating binomial response relations. Biometrika 43(3/4): 461–464.
- Agresti, A. Categorical Data Analysis. Wiley (current edition).

3. Unpaired (Independent Samples) t-Test

3.1 Purpose and scope

The unpaired (independent samples) t-test compares the means of a continuous variable between two independent groups (e.g., Control vs Treatment). Sigma reports both the Welch t-test (unequal variances; default for inference) and the Student t-test with pooled variance (equal variances assumed), along with confidence intervals and standardized effect sizes.

3.2 Input data and preprocessing

The analysis requires a group column with exactly two observed (non-missing) levels and one or more numeric parameter columns analyzed independently. For each parameter, observations with missing or non-numeric entries in either the group or the parameter are excluded for that parameter (complete-case evaluation per parameter).

The user assigns group labels Control and Treatment. Throughout the module, the mean difference is defined as:

$$\Delta = \bar{x}_T - \bar{x}_C,$$

where \bar{x}_T and \bar{x}_C are the sample means in Treatment and Control, respectively.

3.3 Descriptive statistics

For each group and parameter, Sigma reports sample size n , mean \bar{x} , sample variance s^2 (denominator $n - 1$), standard deviation s , standard error $SE = s/\sqrt{n}$, and distribution summaries (min/max, and quartiles). Quartiles are computed from sorted values using a standard linear interpolation approach.

3.4 Welch t-test (unequal variances)

The Welch t-test does not assume equal variances between groups. Let s_T^2 and s_C^2 denote the sample variances. The standard error of the mean difference is:

$$SE_W(\Delta) = \sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}}$$

The Welch t-statistic is:

$$t_W = \frac{\Delta}{SE_W(\Delta)}$$

Degrees of freedom are approximated using the Welch–Satterthwaite equation:

$$df_W = \frac{\left(\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}\right)^2}{\frac{\left(\frac{s_T^2}{n_T}\right)^2}{n_T - 1} + \frac{\left(\frac{s_C^2}{n_C}\right)^2}{n_C - 1}}$$

The two-sided p-value is computed from the Student t distribution with df_W degrees of freedom:

$$p = 2(1 - F_{t,df_W}(|t_W|)),$$

where $F_{t,df}$ denotes the t-distribution CDF.

3.5 Student t-test with pooled variance (equal variances assumed)

Under the assumption of equal variances, Sigma additionally computes the pooled-variance t-test. The pooled variance is:

$$s_p^2 = \frac{(n_T - 1)s_T^2 + (n_C - 1)s_C^2}{n_T + n_C - 2}, s_p = \sqrt{s_p^2}$$

The standard error of the mean difference is:

$$SE_P(\Delta) = s_p \sqrt{\frac{1}{n_T} + \frac{1}{n_C}}$$

and the pooled t-statistic is:

$$t_P = \frac{\Delta}{SE_P(\Delta)}, df_P = n_T + n_C - 2.$$

The two-sided p-value is computed as $p = 2(1 - F_{t,df_P}(|t_P|))$.

(Interpretation: Welch's test is generally preferred when variances differ; Sigma reports both results for transparency. In the summary output, the Welch two-sided p-value is used as the primary inference quantity.)

3.6 Confidence intervals for the mean difference

Sigma reports a two-sided 95% confidence interval for $\Delta = \bar{x}_T - \bar{x}_C$ using the relevant standard error and critical t-value:

$$CI_{95\%} = \Delta \pm t_{1-\alpha/2, df} \cdot SE(\Delta), \alpha = 0.05,$$

where $df = df_W$ and $SE = SE_W$ for Welch, and $df = df_p$ and $SE = SE_p$ for the pooled-variance test.

3.7 Effect sizes

Sigma reports standardized mean differences based on the pooled standard deviation. Cohen's d is computed as:

$$d = \frac{\Delta}{s_p}$$

Hedges' J applies a small-sample correction:

$$J = 1 - \frac{3}{4df_p - 1}, g = J \cdot d.$$

3.8 Assumptions and limitations

- Independence: observations are assumed independent between groups.
- Scale and distribution: the outcome variable should be continuous (or approximately continuous). Normality within each group supports exact t-based inference; in moderate-to-large samples the tests are relatively robust to mild deviations from normality.
- Variance structure: Welch's test accommodates unequal variances; the pooled-variance test assumes equal variances and can be misleading when this assumption is violated.
- Multiple testing: when multiple parameters are tested, Sigma does not automatically apply multiplicity corrections (e.g., Bonferroni/FDR).

3.9 Warnings, edge cases, and safeguards

- Tests are only reported when both groups have sufficient non-missing observations to estimate variance and standard errors.
- Near-constant data (variance ~ 0) can make t-statistics and standardized effect sizes unstable; such situations may be omitted or flagged.
- The sign of Δ , d , and g depends on the Control/Treatment labeling.

3.10 Numerical implementation and equivalence to R

Computations are performed in JavaScript using double-precision floating point arithmetic (IEEE-754). The t-distribution CDF is evaluated numerically, and critical values are obtained by root-finding to match the desired tail probability. Small numerical differences across software may occur due to rounding and numerical integration.

Conceptually, the procedures correspond to R's `stats::t.test` for independent samples: Welch's test (`var.equal = FALSE`) and the pooled-variance test (`var.equal = TRUE`).

3.11 References

- Student (Gosset). (1908). The probable error of a mean. *Biometrika*.
- Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.).
- Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*.

4. Paired t-Test

4.1 Purpose and scope

The Sigma paired t-test module performs a classical paired-samples (dependent-samples) t-test, mathematically equivalent to common implementations in R (`t.test(..., paired = TRUE)`) and standard statistical software. The test compares two measurements obtained from the same subjects (Time 1 and Time 2) and bases inference on the within-subject differences.

4.2 Input data, pairing modes, and preprocessing

The analysis is defined by one or more pairs of numeric columns representing measurements at Time 1 and Time 2. Sigma supports two pairing modes:

- **Manual pairing:** the user explicitly selects the Time 1 column and Time 2 column for each pair (and assigns a label for reporting).
- **Automatic pairing:** Sigma can automatically detect pairs by matching column names using a user-specified suffix pattern (e.g., `_pre / _post`) or prefix pattern (e.g., `pre_ / post_`) for Time 1 and Time 2. Detected pairs can be reviewed and converted into the manual list for reproducible reporting.

For each pair, rows are included if and only if both Time 1 and Time 2 values are present and can be parsed as numeric. Rows with missing or non-numeric values in either column are excluded for that pair (complete-case evaluation per pair).

4.3 Numeric column acceptance (70% rule)

To reduce accidental misuse of non-numeric columns (e.g., text variables), Sigma applies a conservative numeric acceptance rule. A column is treated as numeric only if at least 70% of inspected non-empty entries can be parsed as finite numbers. For this check, Sigma inspects up to the first 30 non-empty entries in the column. Columns failing this criterion are rejected and the corresponding pair is not analyzed.

(This behavior is stricter than R's default coercion, which attempts numeric conversion and produces NA for non-numeric entries. Sigma's rule is intended as a data-quality safeguard.)

4.4 Definition of within-subject differences

Let $x_{1,i}$ denote the measurement at Time 1 and $x_{2,i}$ the measurement at Time 2 for subject i , for all subjects with valid paired observations $i = 1, \dots, n$. Sigma defines the within-subject difference as:

$$d_i = x_{2,i} - x_{1,i}.$$

All subsequent statistics are computed on the vector of differences $d = (d_1, \dots, d_n)$.

4.5 Descriptive statistics

For Time 1, Time 2, and the differences, Sigma reports:

- sample size n ,
- mean \bar{x} (or \bar{d}),
- sample variance s^2 with denominator $n - 1$,
- standard deviation $s = \sqrt{s^2}$,
- standard error $SE = s/\sqrt{n}$,
- minimum and maximum.

4.6 Paired t-test

Let \bar{d} be the mean of the differences and s_d their sample standard deviation:

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i, s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2, SE_d = \frac{s_d}{\sqrt{n}}$$

The paired t-statistic is:

$$t = \frac{\bar{d}}{SE_d} = \frac{\bar{d}}{s_d/\sqrt{n}}, df = n - 1.$$

The two-sided null hypothesis is $H_0: \mu_d = 0$, where μ_d denotes the population mean of differences.

Sigma computes the two-sided p-value from the Student t distribution:

$$p = 2(1 - F_{t,df}(|t|)),$$

Where $F_{t,df}$ is the t-distribution CDF with df degrees of freedom. In Sigma, $F_{t,df}$ and the critical quantiles are obtained via the JavaScript library jStat (jStat.studentt.cdf, jStat.studentt.inv).

4.7 Confidence interval for the mean difference

Sigma reports a two-sided 95% confidence interval for μ_d using $\alpha = 0.05$ and the 97.5% t-quantile

$$t_{crit} = t_{1-\alpha/2,df}:$$

$$CI_{95\%} = \bar{d} \pm t_{crit} \cdot SE_d.$$

4.8 Effect sizes

Sigma reports standardized effect sizes based on the variability of within-subject differences. Cohen's d_z for paired samples is:

$$d_z = \frac{\bar{d}}{s_d}.$$

A small-sample corrected effect size (Hedges' g) is computed using:

$$J = 1 - \frac{3}{4df - 1}, g = J \cdot d_z.$$

4.9 Assumptions and limitations

- Independence at the subject level: pairs are assumed to come from independent subjects; the test does not model clustering beyond the pairing.
- Approximate normality of differences: exact t-based inference assumes the differences are approximately normally distributed in the population. With increasing sample size, the test is generally robust to moderate deviations.

- Measurement scale: outcomes should be continuous (or approximately continuous) and measured consistently at both time points.

4.10 Warnings, edge cases, and safeguards

Sigma reports paired t-test results only when at least two valid pairs are available and the differences are not constant:

- $n \geq 2$ valid paired observations are required.
- If the variance of differences is zero (i.e., $s_d = 0$), no t-statistic is reported.
- Pairs are rejected if one or both columns fail the numeric acceptance rule (70% rule).

4.11 Numerical implementation and equivalence to R

Computations are performed in JavaScript using double-precision floating point arithmetic (IEEE-754). Conceptually, the procedure corresponds to R's `stats::t.test` with `paired = TRUE`. Minor numerical differences across software may occur due to rounding and floating-point arithmetic.

4.12 References

- Student (Gosset). (1908). The probable error of a mean. *Biometrika*.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.).
- Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*.
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

5. Univariate Logistic Regression

5.1 Purpose and scope

The Sigma univariate logistic regression module fits classical binary logistic regression models with an intercept and one predictor at a time. For each selected predictor, Sigma fits a separate univariate model using maximum likelihood estimation (MLE) via an iteratively reweighted least squares (IRLS) algorithm (Newton–Raphson for the logistic log-likelihood). The approach is mathematically equivalent to standard implementations such as R's `glm(..., family = binomial)` for single-predictor models.

5.2 Data structure, outcome definition, and predictors

The tool expects a binary outcome variable and one or more candidate predictor variables. The user selects:

- a binary outcome column,
- the outcome level treated as the event (coded as 1),
- the outcome level treated as the non-event (coded as 0),
- one or more numeric predictors to be analyzed in separate univariate models.

Internally, the outcome is recoded to $y_i \in \{0,1\}$, where $y_i = 1$ denotes the user-defined event $y_i = 0$ and denotes the non-event. Each predictor x_i is treated as continuous (no automatic categorization or spline/nonlinear terms).

(Rows whose outcome does not match either of the two selected levels are excluded from the analysis for that predictor.)

5.3 Missing data handling and numeric checks

For each predictor, Sigma performs a complete-case analysis on the pair (y_i, x_i) :

- rows with missing or non-numeric outcome values are excluded,
- rows with missing or non-numeric predictor values are excluded,
- all remaining rows with valid (y_i, x_i) are included for that predictor's model.

As a data-quality safeguard, a column is treated as numeric only if at least 70% of inspected non-empty entries can be parsed as finite numbers. Sigma inspects up to the first 30 non-empty entries for this check. Columns failing this criterion are not offered as predictors.

For each predictor, Sigma reports:

- n_{total} : number of rows with valid outcome and predictor,
- n_{events} : number of rows with $y = 1$,
- $n_{non-events}$: number of rows with $y = 0$,
- event rate (%): $100 \cdot n_{events}/n_{total}$.

5.4 Logistic regression model

For each predictor, Sigma fits the binary logistic regression model:

$$Y_i \sim \text{Bernoulli}(p_i), \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i,$$

with fitted probability $p_i = \frac{1}{1+\exp(-(\beta_0+\beta_1 x_i))}$. Here, β_1 is the log-odds change per unit increase in x .

The log-likelihood for a sample of size n is:

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)].$$

The MLEs $(\hat{\beta}_0, \hat{\beta}_1)$ maximize ℓ and solve the corresponding score equations.

5.5 Estimation via IRLS (Newton–Raphson)

Sigma estimates parameters using IRLS with an intercept and one slope parameter. Starting from $\beta_0^{(0)} = 0$ and $\beta_1^{(0)} = 0$, iterations proceed up to a maximum of 25 iterations with a convergence tolerance of 10^{-8} on both parameter updates. At iteration t , define the linear predictor η_i , fitted probability p_i , and weights w_i :

$$\eta_i = \beta_0^{(t)} + \beta_1^{(t)} x_i, p_i = \frac{1}{1 + \exp(-\eta_i)}, w_i = p_i(1 - p_i).$$

To reduce numerical overflow and degenerate weights when $|\eta_i|$ becomes very large, Sigma constrains η_i to a finite range (clamp) of approximately $[-20,20]$, preventing fitted probabilities from becoming exactly 0 or 1.

The score components for $\beta = (\beta_0, \beta_1)$ are:

$$U_0 = \sum_i (y_i - p_i), \quad U_1 = \sum_i (y_i - p_i)x_i.$$

The observed Fisher information matrix for the two-parameter model is:

$$I(\beta) = \begin{pmatrix} I_{00} & I_{01} \\ I_{01} & I_{11} \end{pmatrix}, \quad I_{00} = \sum_i w_i, \quad I_{01} = \sum_i w_i x_i, \quad I_{11} = \sum_i w_i x_i^2.$$

The Newton update solves $I(\beta)\Delta\beta = U(\beta)$, and parameters update as $\beta^{(t+1)} = \beta^{(t)} + \Delta\beta$. Iteration stops when both components of $\Delta\beta$ fall below the tolerance or when the maximum number of iterations is reached.

5.6 Standard errors, Wald test, odds ratio, and confidence interval

After convergence (or the final iteration if convergence is not strict), Sigma recomputes the information matrix at the final estimates. For the slope $\hat{\beta}_1$, the variance is derived from the inverse information matrix. In the 2×2 case, with $\det(I) = I_{00}I_{11} - I_{01}^2$, the inverse element used is:

$$\text{Var}(\hat{\beta}_1) = [I(\hat{\beta})^{-1}]_{11} = \frac{I_{00}}{\det(I)}, \quad SE(\hat{\beta}_1) = \sqrt{\text{Var}(\hat{\beta}_1)}.$$

A Wald z-statistic is computed as $z = \hat{\beta}_1 / SE(\hat{\beta}_1)$, and the two-sided p-value is:

$$p = 2(1 - \Phi(|z|)),$$

Where Φ is the standard normal CDF. Sigma uses a fixed critical value $z_{0.975} = 1.9599639845$ for 95% Wald intervals. The odds ratio and its 95% confidence interval are:

$$OR = \exp(\hat{\beta}_1), \quad CI_{95\%}(OR) = [\exp(\hat{\beta}_1 - z_{0.975}SE), \exp(\hat{\beta}_1 + z_{0.975}SE)].$$

5.7 Warnings and quality safeguards

To improve interpretability in pathological data constellations, Sigma attaches warning codes to each fitted univariate model and displays these in the results table and detailed view.

- **Warning 1 — Complete separation / non-identifiable estimates:** Issued when all observations are events ($n_{non-events} = 0$) or all are non-events ($n_{events} = 0$), or when a binary predictor (exactly two unique predictor values) perfectly separates events from non-events. In these situations, the MLE does not exist in the usual sense and the information matrix becomes singular. Sigma suppresses coefficients, standard errors, confidence intervals, and p-values and reports counts with Warning 1.
- **Warning 2 — Very few events or non-events:** Issued when $n_{events} < 5$ or $n_{non-events} < 5$ in the complete-case subset for the predictor. Estimates are computed, but inference may be unstable and confidence intervals wide.
- **Warning 3 — Quasi-separation / numerical instability:** Indicates numerical difficulties or strong evidence of quasi-complete separation, such as lack of strict IRLS convergence, a near-singular information matrix, extremely large coefficient magnitudes, or extreme odds ratios and confidence interval ratios. Estimates are reported whenever numerically feasible, but should be interpreted with extreme caution. In rare extreme cases, standard errors and confidence intervals may be unavailable.

Models are only fitted when both outcome categories are present (at least one event and one non-event) and the predictor shows non-zero variability in the complete-case subset. Otherwise, the model is not fitted and an appropriate warning is issued.

5.8 Assumptions and limitations

- Independence: observations are assumed independent at the subject level.
- Linearity on the logit scale: the model assumes a linear relationship between the predictor and the log-odds of the event; nonlinearity is not modeled (no splines/polynomials).

- Large-sample inference: p-values and confidence intervals rely on Wald (normal) approximations; with small samples or rare events, Wald inference can be unstable.
- No penalized/bias-reduced methods: the current implementation does not include Firth correction or likelihood-ratio-based confidence intervals.

5.9 Numerical implementation and equivalence to R

Calculations are implemented in JavaScript using IEEE-754 double-precision arithmetic. Matrix operations for the 2×2 information matrix are evaluated analytically. The standard normal CDF is computed using a polynomial/rational approximation suitable for statistical inference in typical clinical datasets.

Conceptually, Sigma corresponds to R's `stats::glm` with family = binomial (logit link). Small numerical differences across software may occur due to rounding, convergence criteria, and implementation details, especially in separation scenarios where Sigma intentionally suppresses non-identifiable estimates.

5.10 References

- Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B.*
- McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). Chapman & Hall/CRC.
- Agresti, A. *Categorical Data Analysis*. Wiley (current edition).
- R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

6. Multivariate Logistic Regression

6.1 Overview

The Sigma multivariate logistic regression module implements a classical binary logistic regression model with one intercept and multiple predictors, mathematically equivalent to `glm(..., family = binomial)` in R and standard implementations in SPSS or Stata for multivariable models. Parameters are estimated by maximum likelihood using an iteratively reweighted least squares (IRLS) / Newton–Raphson algorithm in IEEE-754 double precision.

In contrast to the univariate module, all selected predictors enter the same model simultaneously. Predictors with zero variance in the analysis sample are detected, reported, and excluded from the fitted model.

6.2 Data Structure, Outcome Definition and Predictors

The tool requires a binary outcome and at least one numeric predictor. The user specifies:

- a binary outcome column (e.g. 0 = non-event, 1 = event),
- which outcome value is treated as the “event” category,
- which outcome value is treated as the “non-event” category,
- one or more numeric predictor columns to be included together in a single multivariate model.

Internally, the outcome is recoded to $y_i \in \{0, 1\}$, with $y_i = 1$ representing the user-defined event and $y_i = 0$ the non-event. Predictors are treated as continuous without automatic categorization, transformation, or interaction terms.

6.3 Missing Data, Numeric Checks and Listwise Deletion

Sigma applies listwise deletion for the multivariate model: a row is included only if the outcome and all selected predictors are present and numeric. Rows outside the chosen outcome levels are ignored.

Numeric columns are identified conservatively: a column is considered numeric only if at least 70% of up to the first 30 non-empty values can be parsed as numbers (using the selected decimal separator setting). Non-numeric columns are not offered as predictors.

After listwise deletion, let n be the number of complete cases used in the model and k the number of predictors actually fitted (excluding the intercept and excluding predictors dropped due to zero variance). The module reports n , number of events n_{events} , non-events $n_{\text{non-events}}$, and the event rate:

$$\text{Event rate (\%)} = 100 \cdot n_{\text{events}} / n.$$

The events-per-variable ratio (EPV) is computed as:

$$\text{EPV} = n_{\text{events}} / k.$$

6.4 Handling of Constant Predictors

Predictors with zero variance (all values identical) in the analysis sample cannot be estimated separately from the intercept. Such predictors are flagged, excluded from the regression fit, and listed as “dropped predictors” in the model summary. In the results table, these predictors are shown with counts and a warning code indicating that no meaningful coefficient / odds ratio is available.

6.5 Multivariate Logistic Regression Model

For subject i , let Y_i be the binary outcome and $X_i = (x_{i1}, \dots, x_{ik})$ the predictor values. The model is:

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \log(p_i / (1 - p_i)) = \beta_0 + \sum_{j=1}^k \beta_j \cdot x_{ij}$$

where β_j is the log-odds ratio per one-unit increase in predictor X_j , adjusted for all other predictors.

In matrix form, with design matrix X of dimension $n \times (k+1)$ (first column = intercept), parameter vector β , linear predictor $\eta = X\beta$, and probabilities $p_i = 1/(1+\exp(-\eta_i))$, the log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^n \{ y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i) \}.$$

6.6 Estimation via IRLS (Newton–Raphson)

Sigma uses IRLS with a maximum of 25 iterations and a convergence tolerance of approximately 10^{-8} (maximum absolute parameter change). At iteration t :

- $\eta(t) = X \beta(t)$
- $\pi_i(t) = 1/(1+\exp(-\eta_i(t)))$
- $w_i(t) = \pi_i(t)(1-\pi_i(t))$

The score vector and observed Fisher information are computed as:

$$U(\beta(t)) = X^T (y - p(t))$$

$$I(\beta(t)) = X^T W(t) X$$

and the Newton step solves $I \Delta\beta = U$. The linear system is solved by Gauss–Jordan elimination with partial pivoting. Parameters are updated via $\beta(t+1) = \beta(t) + \Delta\beta$.

To prevent numerical overflow and probabilities becoming exactly 0 or 1, the linear predictor is internally clamped to a finite range (approximately $[-20, 20]$). This stabilizes weights and reduces singularity issues in the information matrix.

6.7 Standard Errors, Wald Tests and Confidence Intervals

After convergence (or at the final iteration in borderline cases), the information matrix is recomputed and inverted to obtain the covariance matrix $V = I(\hat{\beta})^{-1}$. Standard errors are:

$$SE(\hat{\beta}_j) = \sqrt{V_{jj}}$$

Wald statistics and two-sided p-values are computed as:

$$z_j = \hat{\beta}_j / SE(\hat{\beta}_j)$$

$$p_j = 2 \cdot (1 - \Phi(|z_j|))$$

where Φ is evaluated using a numerical approximation of the standard normal CDF (sufficient accuracy for routine clinical inference).

Odds ratios and 95% Wald confidence intervals (for predictors; intercept excluded) are:

$$\text{OR}_j = \exp(\hat{\beta}_j)$$

$$\text{CI}_{\text{lower}}(\text{OR}_j) = \exp(\hat{\beta}_j - z_{0.975} \cdot \text{SE}(\hat{\beta}_j))$$

$$\text{CI}_{\text{upper}}(\text{OR}_j) = \exp(\hat{\beta}_j + z_{0.975} \cdot \text{SE}(\hat{\beta}_j))$$

with $z_{0.975} \approx 1.96$. The intercept is used for model fitting and diagnostics but its OR is not displayed.

6.8 Model-Level Diagnostics and Warnings

The model summary reports outcome labels, n, events/non-events, event rate, predictors fitted/selected, EPV, convergence status, and predictors dropped due to zero variance.

Model-level warnings include:

- Warning A: very few events or non-events (e.g. <5) → estimates may be unstable.
- Warning B: low EPV (EPV < 5 high risk; $5 \leq \text{EPV} < 10$ caution).
- Warning C: lack of convergence within the iteration limit → inference may be unreliable.
- Warning D: indications of separation or strong multicollinearity (e.g. near-singular information matrix, failure to solve/invert, extreme coefficient/OR/CI behavior) → OR/CI may be inflated or undefined.

6.9 Predictor-Level Warning Codes

- Warning 1: no meaningful estimate (e.g. constant predictor or non-identifiable coefficient) → $\beta/\text{SE}/\text{OR}/\text{CI}/p$ suppressed.
- Warning 2: very few events or non-events (<5) → unstable estimates, wide CIs likely.
- Warning 3: strong or quasi-separation / numerical instability → estimates may be inflated; interpret with caution.

6.10 Assumptions and Limitations

- Independent observations at the subject level.
- Correct logit link specification; no automatic modeling of non-linearity (no splines, fractional polynomials) and no interactions.
- Wald inference relies on large-sample approximations; for small samples/rare events results may be unstable.
- No penalized/bias-reduced methods (e.g. Firth), no exact logistic regression, and no likelihood-ratio-based intervals.

6.11 Equivalence to R and numerical notes

For well-behaved datasets (adequate events, no extreme separation, no near-singular design), results are expected to agree closely with R's `glm(..., family = binomial)` up to rounding differences. Differences may occur in edge cases because Sigma explicitly flags instability and may leave quantities undefined (e.g. when the information matrix cannot be inverted).

7. Propensity Score Matching (PSM)

7.1 Purpose and scope

The Sigma Propensity Score Matching (PSM) module implements a streamlined propensity score matching workflow for observational comparisons of a binary exposure or treatment T (treated vs. control). Its goal is to reduce confounding by constructing a matched sample with improved balance in measured baseline covariates. The module is intended primarily as a design-stage tool: it estimates propensity scores, performs matching under the implemented fixed settings, and reports overlap and covariate balance diagnostics. Outcome analyses are conceptually performed after matching using methods appropriate for paired matched data.

(Propensity score methods rely on the assumptions of (i) conditional exchangeability / no unmeasured confounding given the selected covariates, and (ii) overlap / positivity, meaning that subjects with similar covariate patterns have a non-zero probability of belonging to either group. These assumptions cannot be verified from observed data alone; the reported diagnostics evaluate only measured covariates and empirical overlap.)

7.2 Notation and estimand

Let $T \in \{0,1\}$ denote treatment assignment, with $T=1$ for treated and $T=0$ for controls, and let X denote the vector of measured baseline covariates. The propensity score is defined as:

$$e(X) = Pr(T = 1 | X).$$

Matching on $e(X)$ aims to create treated and control groups that are comparable with respect to the observed covariates X . In Sigma, the implemented matching procedure is ATT-oriented: treated subjects are the index units, and control subjects are selected as nearest neighbors when suitable matches exist. Consequently, the matched sample is most naturally interpreted as targeting the average treatment effect on the treated (ATT) within the region of empirical overlap.

7.3 Inputs and variable types

- Treatment indicator: the user selects one column and specifies which level represents T=1 (“treated”) and which represents T=0 (“control”). The treatment variable must be binary in the analysis actually performed.
- Covariates: the user selects baseline covariates to be included in the propensity score model. Numeric covariates enter the model as continuous predictors. Binary covariates enter as a single 0/1-coded predictor. Categorical covariates are dummy-coded using reference coding: for a factor with m observed levels, Sigma creates m–1 indicator variables, with the first internal level used as the reference category.

7.4 Missing data handling (as implemented)

In the current Sigma PSM tool, missing data are handled by complete-case analysis only. Rows are excluded before propensity score estimation and matching if any of the following is true:

- the treatment value is missing,
- any selected covariate is missing, or
- a covariate specified as numeric cannot be parsed as a finite numeric value.

(Accordingly, the matched analysis pertains to the subset of observations with complete information on treatment and all selected covariates. Complete-case analysis is transparent and reproducible, but it may reduce sample size and may induce bias if missingness is not completely at random or conditionally ignorable given the retained data.)

7.5 Propensity score model

Sigma estimates propensity scores using a logistic regression model with intercept:

$$T_i \sim \text{Bernoulli}(p_i), \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + X_i^T \beta.$$

The fitted propensity score is:

$$\hat{e}(X_i) = \hat{p}_i = \frac{1}{1 + \exp(-(\hat{\beta}_0 + X_i^\top \hat{\beta}))}.$$

7.6 Estimation (maximum likelihood via IRLS)

Parameters are estimated by maximum likelihood using an iteratively reweighted least squares (IRLS) / Newton–Raphson procedure in double precision. At iteration t , with current parameter vector $\beta^{(t)}$,

$$\eta_i^{(t)} = \beta_0^{(t)} + X_i^\top \beta^{(t)}, p_i^{(t)} = \frac{1}{1 + \exp(-\eta_i^{(t)})}, w_i^{(t)} = p_i^{(t)}(1 - p_i^{(t)}).$$

Using the classical working-response formulation,

$$z_i^{(t)} = \eta_i^{(t)} + \frac{T_i - p_i^{(t)}}{w_i^{(t)}},$$

the updated parameter vector is obtained by solving

$$(X^\top W^{(t)} X) \beta^{(t+1)} = X^\top W^{(t)} z^{(t)}.$$

Iteration stops when parameter changes fall below a numerical tolerance or when the maximum number of iterations is reached. To improve numerical stability in ill-conditioned settings, Sigma enforces a lower bound on the IRLS weights and adds a small ridge term to the diagonal of the weighted normal equations before solving them. Final fitted propensity scores are clamped to the interval $[10^{-6}, 1 - 10^{-6}]$ to avoid exact 0 and 1 values in downstream calculations.

7.7 Distance measure for matching

In the current implementation, matching is performed on the propensity score scale itself. For subject i , the matching score is

$$s_i = \hat{p}_i.$$

The distance between a treated subject i and a control subject j is the absolute difference in fitted propensity scores:

$$d_{ij} = |\hat{p}_i - \hat{p}_j|.$$

7.8 Matching algorithm (nearest neighbor; 1:1; without replacement)

Sigma implements greedy nearest-neighbor matching with a fixed 1:1 ratio and without replacement. Each treated subject is considered as a candidate for matching to at most one control, and each control subject can be used at most once.

Treated subjects are processed in descending order of the chosen matching score (largest propensity score first), with deterministic tie-breaking by row order. For each treated subject, Sigma searches the set of currently available controls and selects the control with the smallest absolute distance d_{ij} , again using deterministic tie-breaking. If no eligible control satisfies the caliper restriction, the treated subject remains unmatched.

Because matching is greedy and without replacement, the final matched sample depends on the sequential allocation of controls. This is methodologically standard for nearest-neighbor matching and is consistent with an ATT-oriented design, but it is not equivalent to optimal global matching.

7.9 Caliper matching

Sigma supports caliper-restricted matching. A candidate treated–control pair is accepted only if

$$d_{ij} \leq c,$$

where c is the caliper threshold. In the current implementation, the user may either specify an absolute caliper directly on the propensity score scale or use a standardized caliper. For the standardized option, Sigma interprets the entered value as a multiple of the empirical standard deviation of the matching distance scale in the full analysis sample. Since the implemented distance scale is the propensity score itself, this becomes

$$c = \gamma \cdot SD(\hat{p}).$$

Thus, with the default standardized setting and caliper value 0.2, Sigma uses $c = 0.2 \times SD(\hat{p})$ in the complete-case analysis sample.

(A smaller caliper improves match closeness but may leave more treated subjects unmatched. Therefore, the practical estimand remains ATT-like but increasingly restricted to the subset of treated subjects lying within empirical overlap.)

7.10 Overlap, common support, and unmatched units

Sigma reports overlap diagnostics on the propensity score scale. These include the pre- and post-matching propensity score ranges in treated and control subjects, their overlap interval, selected quantiles, and the proportion of units lying outside the range of the opposite group. These diagnostics are descriptive and warning-oriented; they do not alter the matching process except through the user-specified caliper.

Unmatched treated units indicate limited overlap and imply that the matched analysis pertains to a subset of treated subjects for whom suitable controls were available under the implemented matching constraints.

7.11 Balance diagnostics (pre- and post-matching)

Sigma reports covariate balance before and after matching. The primary diagnostic is the standardized mean difference (SMD), computed on the design matrix actually used for propensity score estimation, i.e., numeric predictors, binary predictors, and dummy-coded categorical predictors. For categorical covariates, Sigma additionally reports the reference-level contrast reconstructed from the dummy coding so that all observed levels are represented in the balance output.

7.11.1 SMD definition used in Sigma

Sigma uses an ATT-style standardization. For each balance feature Z , let \bar{Z}_T and \bar{Z}_C denote the treated and control means (or proportions for binary indicators), and let $s_{T,\text{pre}}$ denote the standard deviation of Z in the treated group before matching. Then Sigma reports

$$\text{SMD} = \frac{\bar{Z}_T - \bar{Z}_C}{s_{T,\text{pre}}}$$

The denominator $s_{T,\text{pre}}$ is held constant for the pre- and post-matching SMD of the same feature. This yields an ATT-oriented balance metric and makes post-matching changes directly interpretable relative to the original variability in the treated group.

7.11.2 Pre- and post-matching samples

Pre-matching balance is computed in the complete-case analysis sample used for propensity score estimation. Post-matching balance is computed in the matched sample. Because the current implementation uses 1:1 matching without replacement, each matched treated and each matched control contributes weight 1 in the post-matching balance calculations.

7.11.3 Additional descriptive balance metric

Sigma also reports z-diff as an additional descriptive measure of separation between treated and control means or proportions. This quantity depends on the effective sample size and should be interpreted as a descriptive balance diagnostic rather than as a formal hypothesis test.

7.12 Interpreting balance

There is no universal cutoff, but in many applied settings absolute SMD values below approximately 0.1 are considered compatible with good covariate balance. Balance should be assessed across all included covariates rather than relying on a single summary value alone.

7.13 Outputs and exports

Sigma reports:

- the number of rows retained after complete-case preprocessing,
- propensity score model diagnostics (including convergence information and propensity score range),

- matching diagnostics (matched and unmatched treated/control counts, distance scale, caliper, effective caliper),
- overlap / common-support diagnostics on the propensity score scale,
- covariate balance tables with pre- and post-matching SMD and z-diff values.

The module can additionally export the matched cohort and the match pairs, including metadata for downstream analyses. For the current 1:1 without-replacement design, Sigma also supports paired wide-format export.

7.14 After matching: outcome analysis (principles)

Because the current implementation produces 1:1 matched pairs without replacement, downstream outcome analyses should generally account for the paired design. Depending on the outcome type, this may include paired analyses, conditional models, or regression methods that explicitly respect the matching structure. Residual covariate adjustment after matching may be considered when small residual imbalances remain.

(Sigma treats matching primarily as a design-stage procedure and emphasizes transparent reporting of overlap and balance. Correct outcome analysis remains the responsibility of the analyst and depends on the endpoint and estimand of interest.)

7.15 Assumptions and limitations

- No unmeasured confounding: validity depends on whether all important confounders were measured and included in the propensity score model.
- Overlap: limited overlap can leave treated subjects unmatched and changes the target population to treated subjects within the overlap region.
- Model specification: propensity scores are estimated by logistic regression. In propensity score work, the key criterion is achieved covariate balance rather than predictive discrimination alone.

- Complete-case analysis: excluding incomplete observations may reduce sample size and may bias results if missingness is informative.
- Greedy matching: nearest-neighbor matching without replacement is deterministic and practical, but not globally optimal.
- Inference after matching: valid outcome analyses should account for the matched-pair structure rather than treating the matched sample as an ordinary independent sample.

7.16 References

- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983.
- Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 1985.
- Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical Science*. 2010.
- Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*. 2011.
- Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*. 2009.

8. Competing Risk Analysis — CIFs, Gray Test, and Fine & Gray Regression

8.1 Purpose and scope

The Sigma competing risks module provides (i) non-parametric estimation of cumulative incidence functions (CIFs) for each cause with pointwise confidence intervals, (ii) an unadjusted two-group comparison of CIFs using Gray's test ($p = 0$) implemented to match `cmprsk::cuminc`, and (iii) regression modeling of covariate effects on the CIF via the Fine & Gray proportional subdistribution hazards model (effects reported as subdistribution hazard ratios, SHR).

(In competing risks settings, multiple mutually exclusive event types can occur; occurrence of one type prevents occurrence of the others. CIFs quantify absolute risk for a specific cause over time. Gray's test compares CIF curves between groups. Fine & Gray regression models covariate effects on the CIF through the subdistribution hazard.)

8.2 Data structure and preprocessing

The input must contain at least:

- A time variable $time_i$ (numeric, non-negative), representing follow-up time for individual i .
- A status variable $status_i$ with $status = 0$ indicating censoring and positive integers $status \in \{1, 2, \dots\}$ indicating mutually exclusive causes of failure.
- Optionally, a grouping variable with exactly two levels (e.g. "Control" and "Treatment") for Gray's test and group-specific CIFs.
- Optionally, additional covariates for Fine & Gray regression (time-fixed in the current implementation).

Records with missing/non-numeric time, missing status, or invalid time ($time < 0$) are excluded (complete-case analysis). The censoring code is fixed to 0. All strictly positive status values are treated as distinct causes. If a group variable is provided, group labels are internally re-indexed while preserving the user's original labels in outputs. For regression, observations must additionally have non-missing finite covariate values (complete-case with respect to covariates).

8.3 Event-time grid and basic risk set quantities

Let $t_1 < t_2 < \dots < t_m$ be the ordered distinct times at which at least one event (any cause > 0) occurs. For each event time t_j , define:

- n_j : number at risk just before t_j ,
- d_{rj} : number of events from cause r at t_j ,
- $d_j = \sum_r d_{rj}$: total events (all causes) at t_j .

(Ties (multiple events at the same time) are handled by aggregating counts at each distinct time. Individuals with time = t_j are considered at risk just prior to t_j and are removed from the risk set after processing the tied block.)

8.4 Overall survival for “any event”

Sigma computes the Kaplan–Meier survival function for the composite endpoint “any event” (any status > 0), overall and (if grouping is present) within each group. With $S(t_0)=1$, the product-limit estimator on the event-time grid is:

$$\hat{S}(t_j) = \hat{S}(t_{j-1}) \left(1 - \frac{d_j}{n_j}\right), \hat{S}(t_j-) = \hat{S}(t_{j-1}).$$

8.5 Cumulative incidence functions (CIFs)

For cause r , the CIF is the absolute risk of failing from cause r by time t :

$$I_r(t) = Pr(T \leq t, R = r).$$

Sigma uses the standard non-parametric Aalen–Johansen–type estimator on the event-time grid:

$$\hat{I}_r(t_\ell) = \sum_{j:t_j \leq t_\ell} \hat{S}(t_j-) \frac{d_{rj}}{n_j}.$$

(The estimators satisfy $\sum_{r=1}^k \hat{I}_r(t) = 1 - \hat{S}(t)$, i.e. the sum of cause-specific cumulative incidence equals one minus survival for “any event”.)

8.6 Variance and pointwise confidence intervals for CIFs

Sigma reports pointwise standard errors and confidence intervals for each $\hat{I}_r(t)$. Standard errors are computed from a closed-form, large-sample variance estimator for the non-parametric CIF (Aalen–Johansen/Greenwood-type), evaluated on the event-time grid and computed as a sum of time-specific contributions. Small negative values due to rounding are truncated to 0. Standard errors are $SE = \sqrt{\text{Var}}$.

(Note: Sigma reports pointwise intervals at each time point. These are not simultaneous confidence bands.)

Pointwise confidence intervals are constructed using a log(-log) transformation applied directly to the

CIF estimate. For a fixed cause r and time t , let $\hat{I} = \hat{I}_r(t)$ and $SE = \sqrt{\text{Var}(\hat{I})}$. If $0 < \hat{I} < 1$, define:

$$\eta = \ln(-\ln(\hat{I})).$$

Using the delta method, the standard error on the transformed scale is approximated by:

$$SE_\eta \approx \left| \frac{1}{\ln(\hat{I})} \right| SE = \frac{SE}{-\ln(\hat{I})} \text{ (since } \ln(\hat{I}) < 0 \text{)}.$$

With confidence level $1 - \alpha$ (default 95%), let $z = \Phi^{-1}(1 - \alpha/2)$. The CI is computed by forming $\eta_L = \eta - zSE_\eta$ and $\eta_U = \eta + zSE_\eta$ and back-transforming. An equivalent closed form used operationally is:

$$\hat{I}_L = \hat{I}^{\exp(-z \frac{SE}{\ln(\hat{I})})}, \hat{I}_U = \hat{I}^{\exp(+z \frac{SE}{\ln(\hat{I})})}.$$

If \hat{I} is at the boundary (0 or 1) so the transform is undefined, Sigma falls back to a normal approximation on the original scale:

$$\hat{I} \pm zSE,$$

with truncation to [0,1].

8.7 Gray's test for group differences in CIFs (two groups; $\rho = 0$)

If a two-level grouping variable is provided, Sigma computes group-specific CIFs and performs Gray's test for each cause to compare CIF curves between the two groups (unadjusted comparison).

Null hypothesis

For a given cause j , Gray's test assesses:

$$H_0: I_{j,1}(t) = I_{j,2}(t) \text{ for all } t,$$

Implementation aligned to `cmprsk::cuminc`

Sigma implements Gray's test in the same computational form used by the R package `cmprsk` (`cuminc`), i.e. the original algorithm as implemented in the Fortran routines commonly referred to as `crstm/crst`. The test is evaluated on tied event-time blocks and uses the recursive construction of the score and its variance that accounts for (i) the cause of interest and (ii) competing-event contributions, including tie corrections when multiple events occur at the same time.

The weight parameter is fixed to the unweighted case ($\rho = 0$), corresponding to a constant weight function. The resulting test statistic for each cause is reported as a one-degree-of-freedom chi-square:

$$\chi^2 = \frac{U^2}{\hat{\text{Var}}(U)} \sim \chi^2(1) \text{ (asymptotically under } H_0 \text{)}.$$

Scope note: Gray's test in Sigma is restricted to exactly two groups and the unweighted case ($\rho = 0$). Like most large-sample tests, the χ^2 approximation can be unreliable with very small numbers of events for the cause of interest.

8.8 Fine & Gray regression (subdistribution hazard model)

In addition to unadjusted CIF estimation and Gray's test, Sigma implements the Fine & Gray proportional subdistribution hazards model for a chosen cause of interest j . The model targets covariate effects on the CIF via the subdistribution hazard.

Subdistribution hazard and model

$$\lambda_j^*(t | Z) = \lambda_{j0}^*(t) \exp(\beta^T Z),$$

where Z is the covariate vector and $\exp(\beta)$ are subdistribution hazard ratios (SHR).

IPCW for right censoring (with optional censoring strata)

Fine & Gray estimation handles right censoring using inverse probability of censoring weights (IPCW) based on the Kaplan–Meier estimator of the censoring survival function $\hat{G}_C(t)$, constructed by treating censoring as the event and all observed failures as censored in the censoring model. Optionally, $G_C(t)$ can be estimated separately within censoring strata (cengroup), analogous to `cmprsk::crr(cengroup=...)`.

$$w_i(t) = r_i(t) \frac{\hat{G}_C(t)}{\hat{G}_C(X_i \wedge t)}, r_i(t) = \mathbf{1}(C_i \geq (T_i \wedge t)), X_i = \min(T_i, C_i).$$

(Practical conventions are used to avoid division by zero (e.g., truncating/clamping \hat{G}_C away from 0). The validity of IPCW requires non-informative (independent) censoring within the modeling framework.)

Estimation and inference

Sigma solves the weighted estimating equation numerically by Newton–Raphson. After convergence, Sigma reports $\hat{\beta}$, robust (sandwich) standard errors by default, Wald z statistics and p-values, and SHR with 95% Wald confidence intervals.

8.9 Outputs and exports

- Cause-specific CIF tables (overall and, if group provided, by group) on the observed event-time grid,
- Point estimates, standard errors, and pointwise confidence intervals for CIFs,
- Gray’s test statistic and p-value per cause (two groups; $\rho = 0$, aligned to `cmprsk::cuminc`),
- Fine & Gray regression results for the selected cause of interest (SHR, Wald test, CI; robust SE by default),

- Optional exports of CIF tables and model summaries for documentation and downstream analyses.

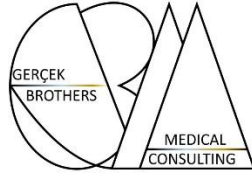
8.10 Assumptions and limitations

- Competing events are assumed mutually exclusive and correctly coded (one cause per subject or censoring).
- Right censoring is assumed independent/non-informative; Fine & Gray uses IPCW based on this assumption.
- Pointwise CIF confidence intervals are provided (not simultaneous bands).
- Gray's test is implemented only for exactly two groups and only for the unweighted case ($p = 0$).
- Fine & Gray regression assumes proportional subdistribution hazards; diagnostics/time-varying effects are not included.
- Fine & Gray regression supports time-fixed covariates only.

8.11 References

- Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics*. 1988.
- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *JASA*. 1999.
- Choudhury JB. Non-parametric confidence interval estimation for competing risks analysis. *Statistics in Medicine*. 2002.
- Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. *Bone Marrow Transplantation*. 2007.
- cmprsk R package documentation (cuminc, crr) as methodological reference for unadjusted CIF/Gray output and Fine & Gray regression reporting.

Methodology for Sigma Statistics from GBM Consulting (07th April 2026)



GBM Consulting
Dr. med. Mustafa Gerçek
Jägerstraße 101A
47228 Duisburg
Deutschland