

From Estimands to Inference: A Practical Tutorial for Robust Statistical Analysis in Scientific Papers

Simone Aonzo, EURECOM

April 15, 2026

Abstract

Modern scientific practice increasingly emphasizes estimation, uncertainty quantification, and transparent reporting over rigid hypothesis testing workflows. This tutorial provides a principled yet practical introduction to core statistical concepts—ranging from basic data summaries to inferential procedures—tailored for researchers conducting standard analyses (two-group comparisons and correlation). We adopt an *estimand-first* perspective, clarifying what is being estimated before selecting methods, and demonstrate how this perspective leads to more robust and interpretable results. The exposition is aligned with a lightweight Python implementation (`inferential_stats.py`); worked examples are implemented in `examples.py` and cross-referenced throughout the paper with stable identifiers (e.g., **EX-RT-01**) that also appear in the script output. It is intended to serve both as a learning resource and as a methodological reference for writing scientifically defensible papers.

Contents

1	Introduction	2
2	Data, Variables, and Assumptions	3
2.1	Independence	4
2.2	Types of Variables	4
3	Descriptive Statistics	4
3.1	Mean vs. Average	5
3.2	Scale	5

4	Statistical Inference	6
4.1	Estimands	6
4.2	Confidence Intervals	7
5	Two-Group Comparisons	7
5.1	Mean Difference	7
5.1.1	Welch's t-test	7
5.1.2	Student's t-test	8
5.2	Effect Size: Hedges' g	8
5.3	Stochastic Dominance	9
5.3.1	Mann–Whitney U Test	9
5.3.2	Cliff's Delta	9
6	Correlation Analysis	10
6.1	Pearson Correlation	10
6.2	Spearman Correlation	11
6.3	Confidence Intervals	11
7	Bootstrap Methods	12
8	Diagnostics and Model Checking	12
9	Practical Workflow	13
10	Reporting Results	14
11	Choosing a Method: Worked Scenarios	15
11.1	Scenario 1: Continuous outcome, two independent groups . .	15
11.2	Scenario 2: Ordinal outcome, two independent groups	15
11.3	Scenario 3: Association between two continuous measurements	15
11.4	Scenario 4: Same question, different estimand	16
12	Conclusion	16

1 Introduction

Statistical analysis in scientific papers is often presented as a sequence of hypothesis tests. However, this perspective can obscure the fundamental goal: *estimating quantities of scientific interest and quantifying uncertainty*. This tutorial reframes standard analyses around this goal and provides a structured path from basic concepts to practical implementation.

The intended audience is a reader who has seen basic statistical terms before but may not yet feel comfortable deciding which method answers which scientific question. For that reason, the paper emphasizes the meaning of each concept before introducing formulas. A useful habit is to ask, at every step: “What quantity am I trying to learn about, and what would a good summary of uncertainty look like?”

We focus on:

- Independent two-group comparisons
- Correlation analysis

These cover a large fraction of empirical studies across disciplines. For example, a researcher may want to compare a treatment group with a control group, or may want to ask whether two measured variables tend to increase and decrease together. Although the mathematical tools differ, the underlying logic is similar: define the quantity of interest, estimate it from data, and report how uncertain the estimate is.

To keep the discussion concrete, every worked example in this paper is implemented in the companion script `examples.py`. Running that script reproduces the datasets and numerical results quoted below.

Code cross-references (PDF ↔ Python). Each worked example below is tagged with a stable identifier such as `EX-RT-01` and a code slug such as `reaction_time_mean_difference`. The same identifier and slug appear in the output of `python examples.py` (and in `python examples.py --json`), so the reader can reliably jump from the PDF to the corresponding code and back.

2 Data, Variables, and Assumptions

We assume data consist of independent observations (e.g., measurements from different individuals) $X_1, \dots, X_n$.

Before discussing methods, it helps to distinguish four closely related ideas:

- A **population** is the larger collection of cases we care about.
- A **sample** is the set of cases we actually observed.
- A **variable** is a measured characteristic, such as age, or reaction time.

- An **observation** is one recorded value of a variable for one unit.

Statistics turns a finite sample into a statement about a population. Because a sample is only one possible draw from the population, uncertainty is unavoidable. Much of inferential statistics is about describing that uncertainty honestly.

2.1 Independence

Independence is the most critical assumption: observations must not influence each other. Intuitively, knowing one observation should not allow us to predict another simply because of the way the data were collected.

For example, if each participant contributes one measurement, independence may be plausible. If the same participant is measured multiple times, the observations are typically related, and methods for repeated measures are needed instead. Ignoring such dependence usually makes uncertainty appear smaller than it really is.

2.2 Types of Variables

We consider:

- Continuous variables (measurements)
- Ordinal variables (rank-like data)

A **continuous** variable can, at least in principle, vary smoothly on a numerical scale: height, blood pressure, and task completion time are common examples. Differences and averages of continuous variables are often directly meaningful.

An **ordinal** variable records order without guaranteeing equal spacing between values. Survey ratings such as “strongly disagree” to “strongly agree” are the standard example. With ordinal data, it is usually safer to reason about ranks or ordering rather than precise arithmetic differences.

The type of variable matters because it affects which summaries and inferential procedures are scientifically interpretable.

3 Descriptive Statistics

Descriptive statistics summarize observed data without generalizing beyond the sample. It characterizes aggregates; they do not reliably describe individual observations, especially in the presence of variability or heavy-tailed distributions.

This section is about describing what was seen, not yet about drawing conclusions about a wider population. Descriptive summaries are still essential because they reveal the basic shape, center, and spread of the data before any formal inference is attempted.

3.1 Mean vs. Average

In statistical writing, it is important to distinguish between the terms *mean* and *average*. The **mean** typically refers to the arithmetic mean, a precisely defined quantity given by:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (1)$$

By contrast, the term **average** is informal and may refer to different measures of central tendency, including the mean, median, or mode. Because of this ambiguity, scientific reporting should favor the term *mean* when referring specifically to the arithmetic average. This ensures clarity and avoids misinterpretation, particularly in skewed or heavy-tailed distributions where different “averages” may differ substantially.

For a novice reader, the key point is that the mean is a balance point of the data. It uses every observation, which makes it informative but also sensitive to unusually large or small values. When a few observations are extreme, the mean can move noticeably even if most of the sample stays the same.

3.2 Scale

The sample standard deviation is:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2)$$

The standard deviation measures how spread out the observations are around the mean. If most values cluster tightly near $\bar{x}$, then s is small. If values are widely dispersed, then s is large.

It is useful to distinguish standard deviation from the *standard error*. The standard deviation describes variability among individual observations. The standard error describes uncertainty in an estimated quantity, such as the sample mean. Beginners often confuse these because both are measures of variation, but they answer different questions.

4 Statistical Inference

Statistical inference uses sample data to estimate population quantities.

The central move in inference is to go beyond “What did this sample look like?” and ask “What does this sample suggest about the population?” Because different samples from the same population would give different answers, inferential methods attach uncertainty statements to estimates.

Four related terms are worth separating:

- An **estimand** is the population quantity we want to know.
- An **estimator** is the rule or formula used to estimate it.
- An **estimate** is the numerical result obtained from the observed sample.
- A **test** is a procedure for assessing compatibility between the data and a hypothesis.

Keeping these terms distinct helps prevent a common mistake: choosing a familiar test first and only later asking what quantity it actually informs.

4.1 Estimands

An **estimand** is the precise quantity of interest. Examples include:

- Difference in means
- Probability of superiority
- Correlation coefficient

This distinction is critical: *methods should be chosen based on the estimand, not vice versa* [Cumming, 2014].

For instance, “Are the groups different?” is often too vague to be scientifically useful. Different methods correspond to different, more precise questions. Are we asking about the difference in average outcome? About whether one group tends to have larger values than the other? About how strongly two variables move together? Once the question is stated precisely, the method selection becomes much clearer.

4.2 Confidence Intervals

A confidence interval (CI) provides a range of plausible values for the estimand.

In practice, a confidence interval combines two ideas: the estimated effect size and the precision of that estimate. A narrow interval suggests the data pin down the estimand fairly well; a wide interval suggests substantial uncertainty.

For a 95% CI, the frequentist interpretation is about the procedure, not about one fixed realized interval. If we repeated the study many times and computed a 95% CI each time, about 95% of those intervals would contain the true estimand. In applied writing, it is common and useful to say that the interval gives a range of values reasonably compatible with the data and the model assumptions.

5 Two-Group Comparisons

Let X and Y denote two independent samples.

This setting appears whenever we compare two conditions, groups, or treatments. The most important first decision is what kind of difference matters scientifically. Sometimes the target is a difference in means; sometimes it is an ordering statement such as whether values from one group tend to exceed values from the other.

5.1 Mean Difference

The estimand is:

$$\Delta = \mathbb{E}[X] - \mathbb{E}[Y] \tag{3}$$

This quantity asks how much larger or smaller the population mean of group X is than the population mean of group Y . If $\Delta > 0$, group X tends to have larger values on average. If $\Delta < 0$, group Y tends to have larger values on average.

The mean difference is especially natural when the variable is continuous and differences on the original measurement scale are substantively meaningful. For example, a difference of 5 milliseconds or 2 exam points may be directly interpretable in a way that a rank-based quantity is not.

5.1.1 Welch's t-test

Welch's t-test [Welch, 1947] estimates Δ without assuming equal variances.

This is often the safer default for comparing means from two independent groups. The reason is simple: real datasets frequently have different variability in the two groups, and Welch’s method is designed to remain valid under that inequality. It still targets the same estimand, the difference in means.

Despite the name “test,” the method is also useful because it naturally leads to a confidence interval for the mean difference. For most scientific reporting, the interval and the estimated difference are at least as important as the p-value.

Degrees of freedom. Degrees of freedom (df) appear in outputs such as $t(24.48) = -8.38$ and are used to determine p -values and confidence intervals. Conceptually, the degrees of freedom represent the amount of independent information available to estimate variability. In simple settings, this is closely related to sample size, but it is reduced by the need to estimate quantities such as group means or variances.

In Welch’s t -test (used throughout this tutorial for mean differences), the degrees of freedom are not necessarily an integer. This is because the method adjusts for unequal variances between groups, producing an *effective* degrees of freedom that reflects both sample sizes and variability. Smaller df correspond to greater uncertainty and heavier tails in the reference distribution, leading to wider confidence intervals and more conservative inference.

For interpretation, the key point is that df are not parameters of scientific interest; they are part of the inferential machinery that determines how uncertainty is quantified. When reporting results, df should be included for completeness (e.g., $t(24.48) = -8.38$), but substantive interpretation should focus on the estimated effect, its confidence interval, and its practical meaning.

5.1.2 Student’s t -test

Student’s t -test assumes equal variances [Student, 1908].

This equal-variance assumption means the two groups are treated as if they have the same underlying spread. When that assumption is wrong, uncertainty can be misestimated. For this reason, many analysts prefer Welch’s approach unless there is a strong substantive reason to pool variances.

Example (EX-RT-01; code slug `reaction_time_mean_difference`): a reaction-time experiment. Suppose a psychologist compares reaction times for participants using interface A versus interface B. In the dataset

from `examples.py`, interface A has $n = 12$ observations with mean 406.92 ms, and interface B has $n = 11$ observations with mean 430.00 ms. Because the scientific question is “How many milliseconds faster is one interface on average?”, the estimand is the difference in means. Welch’s t-test estimates that difference as -23.08 ms (95% CI $[-33.92, -12.25]$), with $t(20.96) = -4.43$ and $p < .001$. The negative sign means interface A is faster on average, and the result is easy to report directly in milliseconds.

5.2 Effect Size: Hedges’ g

Hedges’ g standardizes the mean difference [Hedges, 1981].

Whereas Δ is expressed in the original measurement units, Hedges’ g expresses the group difference relative to the variability in the data. Standardized effect sizes are helpful when raw units are unfamiliar or when comparing results across studies that use different measurement scales.

However, standardization can also make results less intuitive for a novice reader. In many papers, the best practice is to report both: the raw mean difference for interpretability and a standardized effect size for comparability.

5.3 Stochastic Dominance

An alternative estimand is the probability of superiority:

$$\theta = P(X > Y) + \frac{1}{2}P(X = Y) \quad (4)$$

This quantity is often easier to interpret verbally than it first appears. Imagine drawing one observation from group X and one from group Y at random. Then θ is the probability that the draw from X is larger than the draw from Y , counting ties as half a success.

If $\theta = 0.5$, neither group tends to dominate the other. Values above 0.5 suggest group X tends to be larger; values below 0.5 suggest group Y tends to be larger. This estimand is appealing for ordinal data and for situations where rank ordering matters more than mean differences.

5.3.1 Mann–Whitney U Test

This test evaluates stochastic dominance [Mann and Whitney, 1947].

Beginners often hear that the Mann–Whitney test is “the nonparametric version of the t-test.” That shortcut is only partly helpful. The more precise view is that it is a rank-based procedure tied to ordering between groups. It is best understood through the estimand it informs, rather than as a universal replacement for mean-based methods.

5.3.2 Cliff’s Delta

Cliff’s delta is defined as:

$$\delta = 2\theta - 1 \tag{5}$$

Cliff’s delta is a simple transformation of the probability of superiority. Its values range from -1 to 1 :

- $\delta = 0$ means neither group tends to be larger.
- $\delta > 0$ means group X tends to have larger values.
- $\delta < 0$ means group Y tends to have larger values.

Because it is based on pairwise ordering rather than raw distances, Cliff’s delta is especially useful when the data are ordinal or contain outliers that make mean-based summaries less stable.

Example (EX-OR-01; code slug `usability_ratings`): an ordinal usability rating study. Imagine two versions of a mobile app are compared using a 5-point satisfaction scale from “very dissatisfied” to “very satisfied.” In the worked dataset, both versions have $n = 12$ ratings; the newer version has mean rating 4.25 and the older version has mean rating 3.50. These ratings are ordered, but the distance between adjacent categories is not guaranteed to be equal, so a rank-based approach is more defensible than comparing means of the numeric labels directly. The Mann–Whitney analysis estimates a probability of superiority of 0.76 (95% CI [0.57, 0.91]), with $U = 109$ and $p = .022$, while Cliff’s delta is 0.51. In plain language, a randomly chosen rating from the new version exceeds one from the old version about 76% of the time, counting ties as one-half.

6 Correlation Analysis

Correlation addresses a different kind of scientific question from a group comparison. Instead of comparing two samples, we now study whether two variables vary together across observations.

Two warnings are worth stating early. First, correlation does not imply causation: two variables can be associated for many reasons, including common causes or selection effects. Second, a correlation close to zero means lack of *linear* or *monotonic* association depending on the measure used; it does not guarantee the absence of any relationship.

6.1 Pearson Correlation

Pearson’s correlation coefficient is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (6)$$

It measures linear association [Pearson, 1895].

The coefficient r ranges from -1 to 1 . Values near 1 indicate a strong positive linear relationship, values near -1 indicate a strong negative linear relationship, and values near 0 indicate weak linear association.

Pearson correlation is appropriate when the relationship of interest is approximately linear and when unusually extreme observations are not dominating the result. A scatterplot is often the fastest way to judge whether Pearson’s r is a sensible summary.

6.2 Spearman Correlation

Spearman’s ρ is Pearson’s correlation applied to ranks [Spearman, 1904].

This makes Spearman’s correlation sensitive to *monotonic* relationships, meaning relationships that mostly move in one direction even if they are not straight lines. If larger values of one variable tend to go with larger values of the other, Spearman’s ρ will be positive even when the pattern is curved rather than linear.

Because it is rank-based, Spearman’s correlation is often more robust to outliers and more appropriate for ordinal data.

Example (EX-CO-01; code slug `study_time_exam_pearson`): study time and exam performance. Suppose an instructor records weekly study hours and final exam scores for a group of students. In the dataset from `examples.py`, the sample size is $n = 15$. The Pearson correlation is $r = 0.93$ (95% CI $[0.80, 0.98]$, $p < .001$), and the Spearman correlation is $\rho = 0.94$ (95% CI $[0.79, 0.99]$, $p < .001$). Because the two summaries are very similar, this example illustrates a setting where the relationship is increasing and close enough to linear that Pearson’s r is a sensible default, while Spearman’s ρ reaches the same substantive conclusion.

6.3 Confidence Intervals

For Pearson correlation, Fisher’s z transform is used [Fisher, 1921]:

$$z = \tanh^{-1}(r) \quad (7)$$

The purpose of this transformation is not to change the scientific question, but to make uncertainty calculations work better mathematically. Correlations have a bounded scale, so direct normal approximations on the raw r scale can behave poorly, especially when $|r|$ is large. Fisher’s transformation maps the correlation to a scale where approximate normality is more reasonable.

For Spearman correlation, bootstrap methods are often used [Efron and Tibshirani, 1994].

This is common because the sampling distribution of rank-based statistics can be awkward in finite samples. Bootstrap intervals use repeated resampling from the observed data to approximate the amount of uncertainty without requiring a simple closed-form formula.

7 Bootstrap Methods

Bootstrap resampling estimates uncertainty by repeatedly sampling with replacement [Efron, 1979].

The phrase “with replacement” means that after an observation is drawn into a bootstrap sample, it is placed back and may be drawn again. As a result, a bootstrap sample has the same size as the original sample but is not identical to it: some observations appear more than once and some not at all.

The bootstrap is conceptually simple:

1. Treat the observed sample as a stand-in for the population.
2. Resample from it many times.
3. Recompute the statistic of interest on each resample.
4. Use the resulting distribution to estimate standard errors or confidence intervals.

For beginners, the bootstrap is valuable because it links uncertainty to a concrete computational experiment rather than only to abstract probability formulas.

Example (EX-BS-01; code slug `stress_sleep_spearman`): confidence intervals for a hard-to-analyze statistic. Suppose a researcher wants a confidence interval for Spearman’s correlation between stress level and sleep quality in a modest sample. In the companion dataset, $n = 15$ paired

observations give a Spearman correlation of $\rho = -0.96$ with a bootstrap 95% CI of $[-0.99, -0.84]$ and $p < .001$. Rather than relying on a complicated analytic approximation, the researcher can repeatedly resample the observed participant pairs, recompute Spearman's ρ for each resample, and use the resulting empirical distribution to form the interval. This illustrates the practical appeal of bootstrap methods: they provide a flexible route to uncertainty quantification when exact formulas are inconvenient.

8 Diagnostics and Model Checking

Diagnostics should inform interpretation, not dictate method selection.

In practice, diagnostics ask whether the data look compatible with the assumptions that support a method, and whether any observations are so unusual that they deserve closer attention. They are part of understanding the data, not a mechanical gatekeeping ritual.

For two-group comparisons, common checks include plotting both groups, inspecting outliers, and considering whether the variable is so skewed that a mean-based summary becomes hard to interpret. For correlation, a scatterplot is especially important because it can reveal curvature, clusters, or a single influential point that a summary coefficient might hide.

The goal is not perfection. Real data are messy. The goal is to understand whether the numerical summaries are telling a stable and scientifically meaningful story.

9 Practical Workflow

1. Define the estimand
2. Choose method aligned with estimand
3. Compute estimate and CI
4. Report results with interpretation

This workflow is intentionally simple, but it captures a disciplined way of thinking:

- First decide what scientific quantity matters.
- Then select a method that estimates that quantity under assumptions you can defend.

- Next compute both the estimate and its uncertainty.
- Finally explain the result in words that connect back to the original research question.

Many statistical mistakes occur when these steps are reversed, especially when a researcher begins with a familiar test and only afterward tries to interpret whatever number comes out. An estimand-first workflow helps prevent that problem.

The companion `examples.py` script follows this workflow for every example in the paper: it defines the estimand, runs the aligned analysis, and prints the resulting estimates, intervals, and test summaries.

10 Reporting Results

A complete report includes different items to answer different questions about the data and the scientific question.

- The **point estimate** states the best single-number summary from the data.
- The **confidence interval** expresses uncertainty around that estimate.
- The **test statistic and p-value** quantify how incompatible the data are with a specific null model.
- The **effect size** helps judge practical magnitude rather than only statistical detectability.

In a hypothesis test we choose a *test statistic* $T(\mathbf{X})$: a single number computed from the data $\mathbf{X}$ that would tend to be large (or small) when the null hypothesis H_0 is false. After observing $t_{\text{obs}} = T(\mathbf{x}_{\text{obs}})$, the *p-value* is the probability, *assuming H_0 and the rest of the model are true*, of getting a test statistic at least as extreme as what was observed. For a two-sided test this is commonly written as

$$p = \Pr(|T(\mathbf{X})| \geq |t_{\text{obs}}| \mid H_0),$$

where “extreme” is defined with respect to the null distribution of $T(\mathbf{X})$ (for example a t distribution for Welch’s t -test, or the null distribution used for U in a Mann–Whitney test). For a one-sided test, the tail is taken in the direction specified by the alternative hypothesis. The p-value is *not*

the probability that H_0 is true; it is a statement about how surprising the observed data would be *if* H_0 were true.

For novice-friendly scientific writing, it is usually better to lead with the estimate and interval, and only then report the p-value. This keeps attention on scientific magnitude and uncertainty rather than reducing the result to a binary significant/not-significant label.

A clear sentence often has the following structure: report the estimated quantity, give its interval, then provide the associated test result. For example, one might write that the drug group had systolic blood pressure 8.95 mmHg lower on average than the placebo group, with a 95% CI from 6.74 to 11.15 mmHg lower, together with the relevant test statistic and p-value.

11 Choosing a Method: Worked Scenarios

Because beginners often struggle with method selection, so it is helpful to look at complete research scenarios.

11.1 Scenario 1: Continuous outcome, two independent groups

A biomedical researcher compares systolic blood pressure after treatment in a drug group and a placebo group. In `examples.py`, the drug group has $n = 14$ and mean 119.29 mmHg, whereas the placebo group has $n = 13$ and mean 128.23 mmHg. Blood pressure is continuous, and a difference in mean millimeters of mercury is directly meaningful, so the estimand is the difference in means. Welch's t-test estimates the mean difference as -8.95 mmHg (95% CI $[-11.15, -6.74]$), with $t(24.48) = -8.38$ and $p < .001$. This is a good example of a continuous outcome where the raw-unit effect is immediately interpretable.

11.2 Scenario 2: Ordinal outcome, two independent groups

An education researcher compares two teaching methods using a 7-point self-confidence rating collected after the course. In the worked example, both groups have $n = 14$, with mean ratings 5.86 for method A and 4.86 for method B. The scale is ordered, but differences between adjacent ratings may not be meaningfully equal, so a stochastic-dominance perspective is preferable. The Mann-Whitney analysis estimates a probability of superiority of 0.82 (95% CI $[0.67, 0.93]$), with $U = 160$ and $p = .002$, and Cliff's delta is 0.63. This means a randomly chosen student from method A is rated

more confident than one from method B about 82% of the time, counting ties as one-half.

11.3 Scenario 3: Association between two continuous measurements

A physiologist records body temperature and heart rate during exercise sessions. In the companion example, the sample size is $n = 12$, and Pearson correlation gives $r = 0.96$ (95% CI $[0.88, 0.99]$, $p < .001$). If the goal is to summarize whether the variables are linearly associated across sessions, Pearson correlation is suitable because the paired measurements move together in a roughly linear way. If the pattern were monotonic but clearly curved, or if a few sessions were extreme, Spearman correlation would be a more stable summary.

11.4 Scenario 4: Same question, different estimand

Consider again the reaction-time dataset from earlier. One researcher asks, “How many milliseconds faster is interface A than interface B on average?” That question targets a difference in means, and the answer is a mean difference of -23.08 ms (95% CI $[-33.92, -12.25]$). Another asks, “If I randomly pick one user from each interface, how likely is the user on interface A to finish faster?” To answer that question, we switch to a stochastic-dominance framing on the completion-time scale and estimate that interface A is faster about 90% of the time (probability of superiority = 0.90, 95% CI $[0.75, 1.00]$, $U = 119$, $p = .001$). The dataset is identical in both analyses, but the scientifically relevant summary changes because the estimand changes.

Because the outcome is completion time, “A is faster” means $T_A < T_B$, so in code we compute $P(T_B > T_A)$, which is equivalent to the probability that A finishes faster than B (counting ties as one-half).

12 Conclusion

This tutorial provides a principled foundation for statistical analysis in scientific papers, emphasizing clarity of estimands, robustness of methods, and transparency of reporting.

For a novice reader, the most important lesson is that statistics is not mainly about memorizing test names. It is about linking a scientific question to a quantity of interest, using data to estimate that quantity, and describing how certain or uncertain that estimate is. Once that perspective

is clear, individual methods such as Welch’s t-test, Mann–Whitney, Pearson correlation, Spearman correlation, and the bootstrap fit into a much more coherent picture.

References

- Geoff Cumming. The new statistics: Why and how. *Psychological Science*, 25(1):7–29, 2014. doi: 10.1177/0956797613504966.
- Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 1994.
- R. A. Fisher. On the ”probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.
- Larry V Hedges. Distribution theory for glass’s estimator of effect size and related estimators. *journal of Educational Statistics*, 6(2):107–128, 1981.
- Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.
- Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041.
- C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.
- Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.
- B. L. Welch. The generalization of Student’s problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.