Hypothesis testing basics & simple regression

Faculty of DS & AI
Spring semester, 2026

Trong-Nghia Nguyen

Business AI Lab

Content

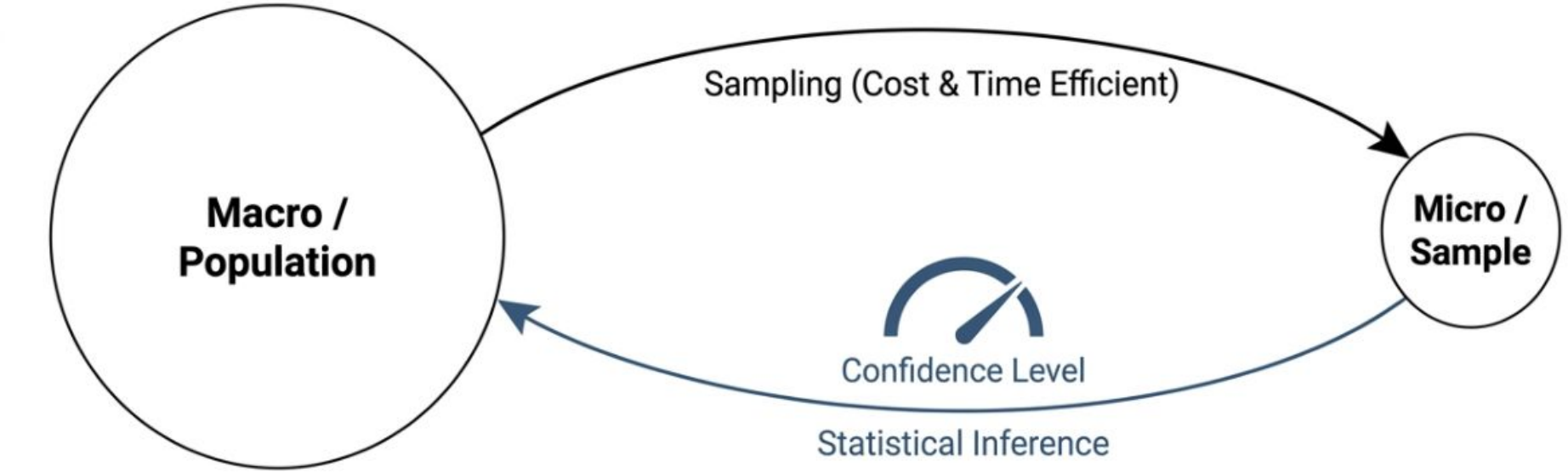
- **Inferential statistics and hypothesis testing**
- **Average testing tools in Excel**
- **Univariate linear regression**
- **Output Summary**

Inferential statistics and hypothesis testing

The Statistical Paradigms Matrix

Descriptive Statistics	Inferential Statistics
Focus: Summarizing the known. **Action:** Organizing and describing characteristics of an existing dataset.	**Focus:** Predicting the unknown. **Action:** Using sample data and probability theory to draw conclusions or forecasts about a larger population with a specific level of confidence.

The Inference Bridge



Inferential statistics and hypothesis testing

Ecosystem of Data

μ	**Population Mean**	**The Population**
	Definition: The complete set of all subjects under study.	
	Business Example: All customers of a national bank.	
$\bar{X}$	**Sample Mean**	**The Sample**
	Definition: A representative group selected from the population.	
	Business Example: 1,000 bank customers participating in a survey.	

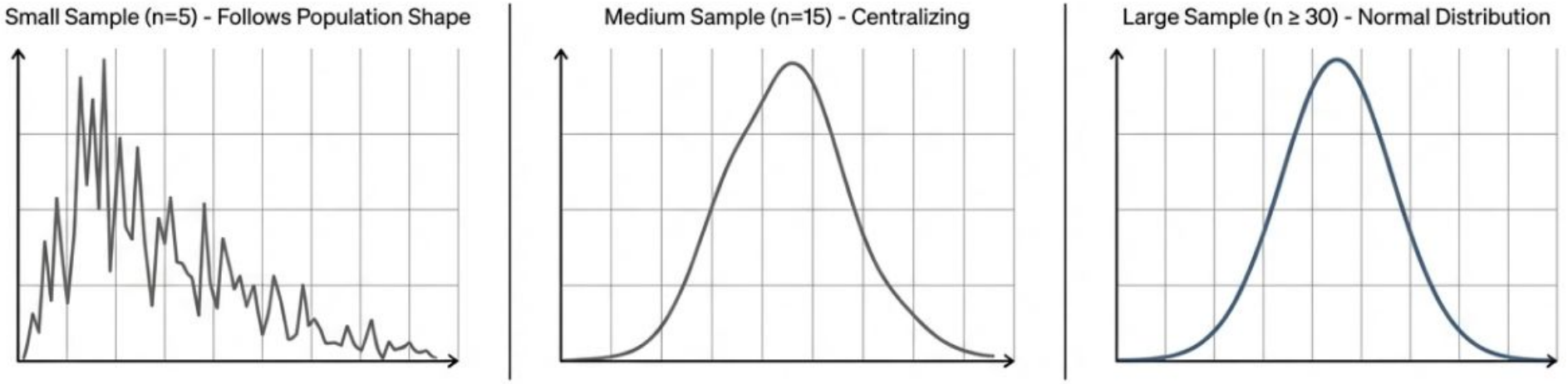
We calculate $\bar{X}$ to estimate μ because μ is practically unobservable.

Inferential statistics and hypothesis testing

The Mathematical Anchor: Central Limit Theorem (CLT)

The CLT is the cornerstone of inferential statistics, enabling parametric tests like Z-tests and T-tests.

CLT Convergence Graphic



As sample size (n) grows, the distribution of sample means ($\bar{X}$) approximates a Normal Distribution, regardless of the original population's shape.

Inferential statistics and hypothesis testing

The Threshold of Normalcy

$$n \geq 30$$

Axiom 1: The Rule of Thumb

In general practice, a sample is considered "sufficiently large" to invoke the Central Limit Theorem when $n \geq 30$.

Axiom 2: Business Significance

This threshold explains why standard statistical tools can be reliably used to analyze highly skewed business variables—such as income distribution, customer spending, or product lifespan—even when the underlying raw data is completely non-normal.

Inferential statistics and hypothesis testing

The Logic of Hypothesis Testing

Hypothesis testing does not prove **absolute truth**. It determines if sample evidence is strong enough to reject a baseline assumption.

The Hypothesis Matrix

Null Hypothesis (H_0)	Alias: "The Status Quo" Definition: The assumption of "no change" or "no difference". Business Example: A new marketing campaign does not increase average revenue (H_0: $\mu_{new} = \mu_{old}$).
Alternative Hypothesis (H_a or H_1)	Alias: "The Challenger" Definition: The claim the analyst is actively trying to find evidence to support. Business Example: The new marketing campaign successfully increases revenue (H_a: $\mu_{new} > \mu_{old}$).

Inferential statistics and hypothesis testing

The Diagnostic Methodology: Setup and Calculation

Step 1: Formulate Hypotheses

Define H_0 and H_a based strictly on the specific business objective.

→

Step 2: Select Significance Level (α)

Define the acceptable risk threshold (commonly set to 0.05 or 5%). This represents the probability of incorrectly rejecting H_0.

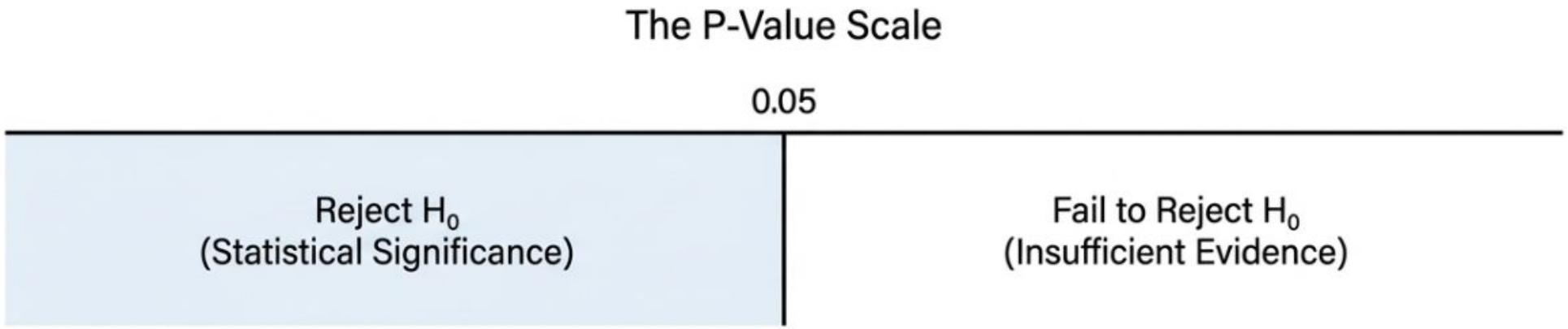
→

Step 3: Collect Data & Compute

Extract the necessary sample statistics: Sample Mean ($\bar{X}$), Standard Deviation (s), and Sample Size (n).

Inferential statistics and hypothesis testing

The Diagnostic Methodology: Evaluation and Decision



Step 4 - The P-Value

The probability of observing the current result (or a more extreme one) assuming the Null Hypothesis (H_0) is entirely true.

Step 5 - The Decision

If $p < \alpha$, the evidence supports the challenger.

If $p \geq \alpha$, the status quo holds.

Inferential statistics and hypothesis testing

The Anatomy of Decision Error

In business, every decision derived from sample data carries an inherent risk of error. A mathematically flawless hypothesis test can still yield a business failure if the sample does not perfectly reflect reality.

Inferential statistics and hypothesis testing

The Ultimate Diagnostic Matrix

Statistical Decision	Objective Reality: H_0 is Actually True	Objective Reality: H_0 is Actually False
Reject H_0	**Type I Error (α)** **False Positive** **Business consequence:** Changing a well-functioning process, leading to wasted investment.	Correct Decision
Fail to Reject H_0	Correct Decision	**Type II Error (β) / False Negative** **Business consequence:** Missing a crucial opportunity for improvement or failing to detect a critical product flaw.

Inferential statistics and hypothesis testing

Context (Manufacturing)

Quality control on a production line. A packaging machine must fill milk cartons with exactly 3000g of product, as committed on the packaging.

Hypothesis Setup

H_0: $\mu = 3000g$ (The machine is calibrated correctly).

H_a: $\mu \neq 3000g$ (The machine is miscalibrated).

Risk of Error

Type I Risk: Stopping a perfectly fine production line for unnecessary repairs (wasted time/money).

Type II Risk: Shipping underweight milk to consumers (reputation damage/legal risk) or overweight milk (lost product margin).

Inferential statistics and hypothesis testing

Context (Marketing)

A/B testing digital assets. Evaluating whether Version A of an advertisement generates a different Click-Through Rate (CTR) than Version B.

Hypothesis Setup

H_0: $CTR_A = CTR_B$ (The new ad performs the same as the old ad).

H_a: $CTR_A \neq CTR_B$ (The new ad drives a genuinely different behavioral response).

Risk of Error

Type I Risk: Paying to deploy a new campaign that actually offers no performance benefit over the legacy creative.

Type II Risk: Discarding a highly effective new ad variation, leaving potential conversion revenue on the table.

Inferential statistics and hypothesis testing

Context (Finance)

Portfolio risk analysis. Auditing a loan portfolio to ensure the bad debt ratio has not exceeded the strictly mandated safety threshold of 3%.

Hypothesis Setup

H_0: $\mu \leq 3\%$ (The portfolio is safe and within acceptable risk limits).

H_a: $\mu > 3\%$ (The portfolio has breached the risk threshold).

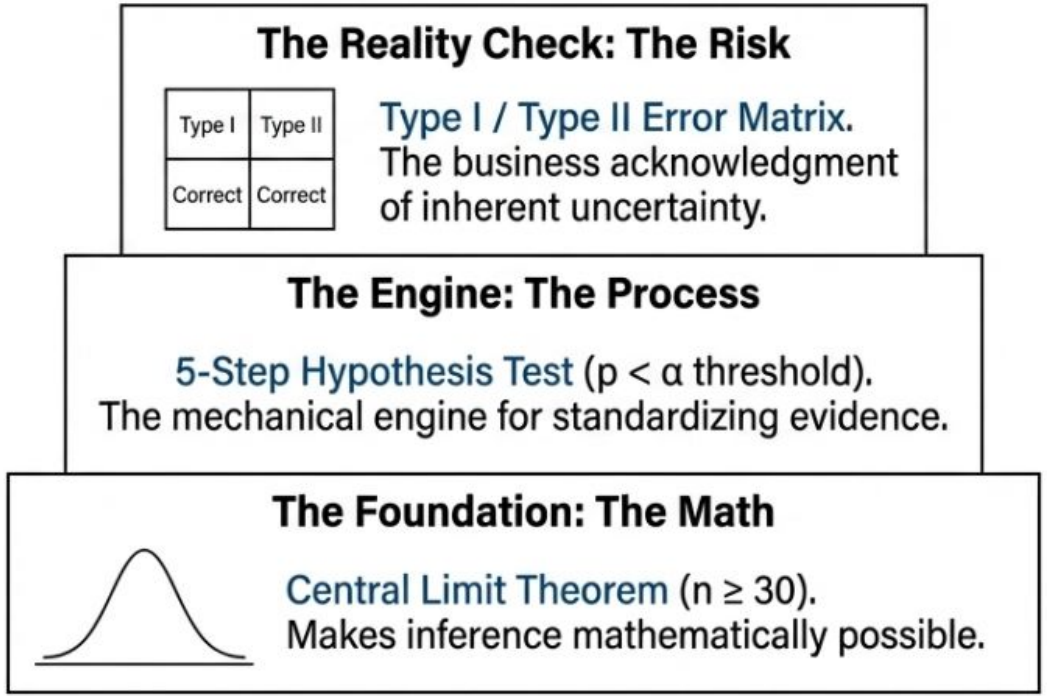
Risk of Error

Type I Risk: Unnecessarily freezing loan issuances, severely restricting the bank's revenue generation.

Type II Risk: Falsely believing the portfolio is safe, potentially leading to catastrophic institutional liquidity crises.

Inferential statistics and hypothesis testing

The Unified Framework of Data-Driven Decisions



Hypothesis testing does not discover absolute truth. It provides a mathematically rigorous, risk-managed architecture for making high-stakes business decisions based on inevitably incomplete information.

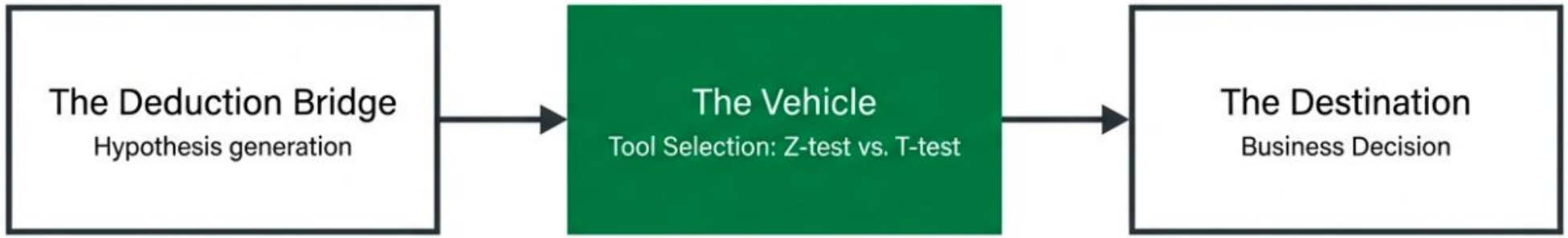
Content

- Inferential statistics and hypothesis testing
- **Average testing tools in Excel**
- Univariate linear regression
- Output Summary

Average testing tools in Excel

Choosing the Right Vehicle for the Deduction Bridge.

Selecting the wrong testing tool in Excel leads to a distorted p-value. A distorted p-value inherently results in fundamentally flawed business decisions, regardless of how robust the initial hypothesis was.

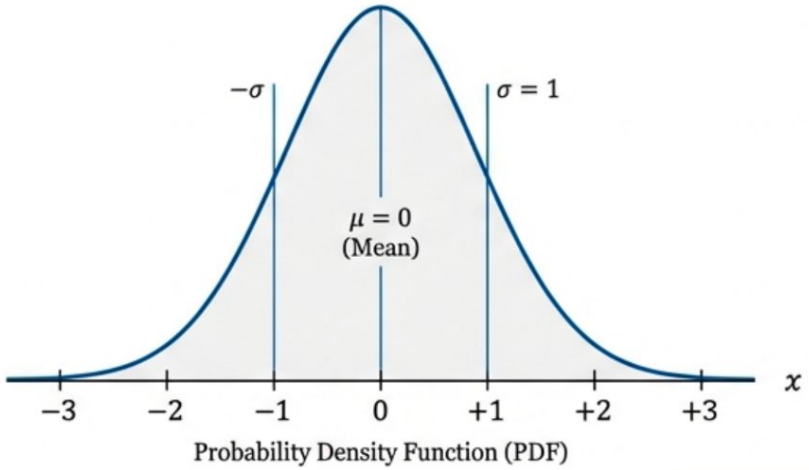


Average testing tools in Excel

Standard Normal Distribution (Z)

- Expected Value (Mean): $\mu = 0$
- Standard Deviation: $\sigma = 1$
- Symmetric, perfect bell-shape
- Total area under the curve = 1

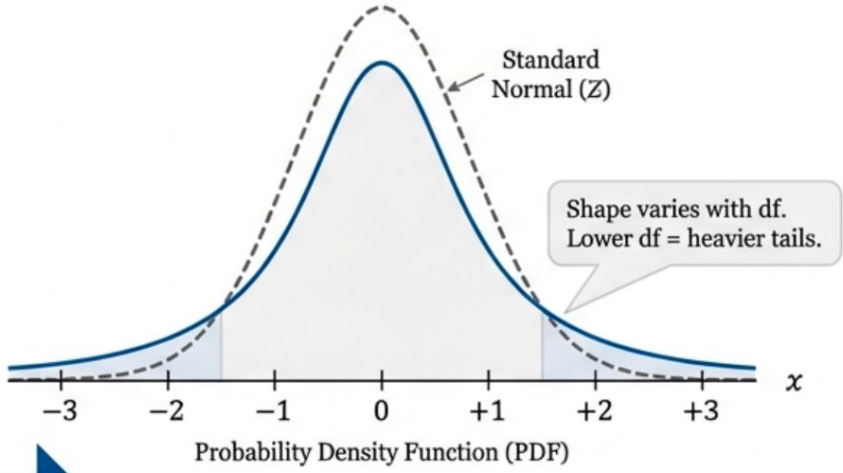
$$\Phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$



Student's *t*-Distribution (t)

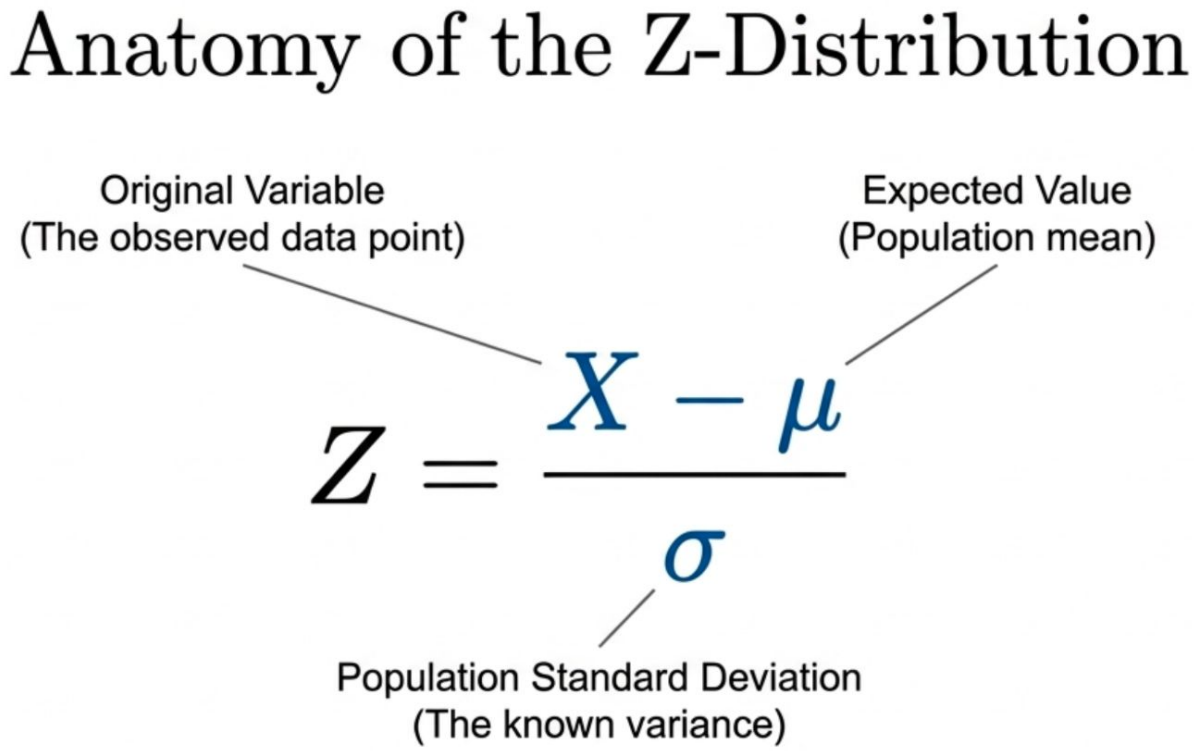
- Used when population standard deviation (σ) is unknown
- Relies instead on the sample standard deviation (s)
- Shape depends on Degrees of Freedom ($df = n - 1$)

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



As $n \to \infty$, $t \to Z$

Average testing tools in Excel

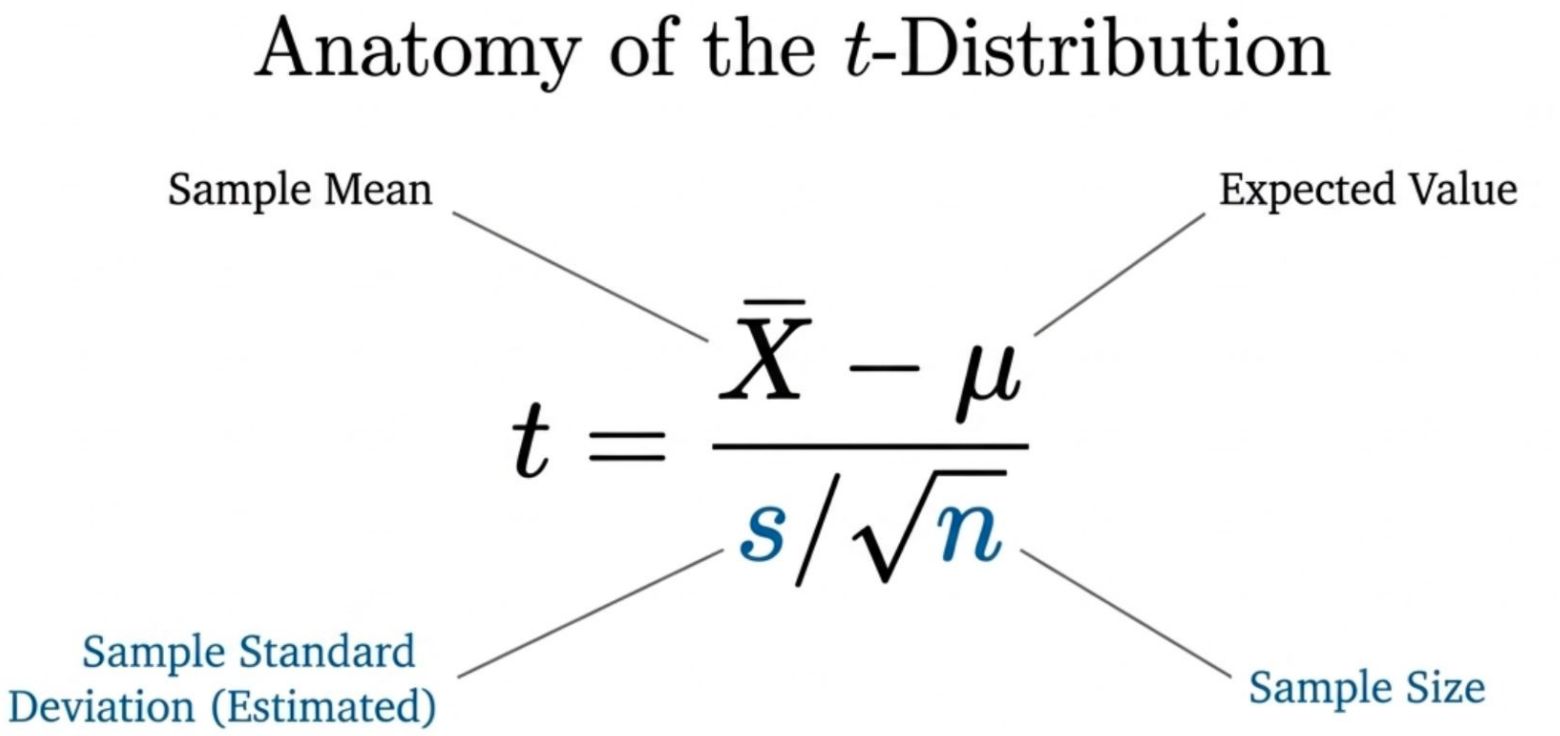


When to Use:

Condition 1: The true population σ is known.

Condition 2: The sample size is large ($n \geq 30$), relying on the Central Limit Theorem.

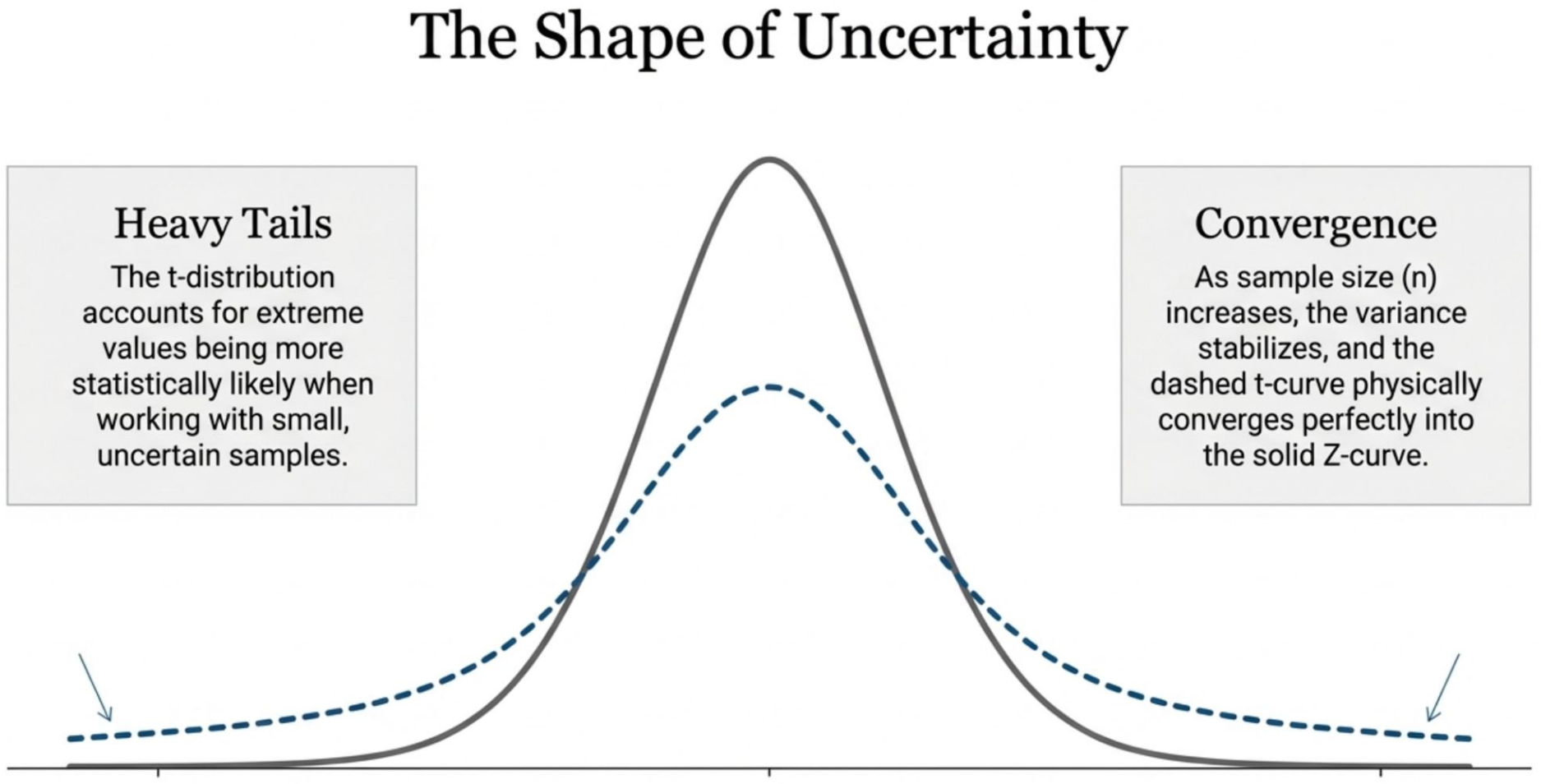
Average testing tools in Excel



Managing Uncertainty:

- The formula reflects the mathematical uncertainty of estimating σ from a limited sample.
- Requires the calculation of Degrees of Freedom ($df = n - 1$) to determine the exact curve shape.

Average testing tools in Excel



Average testing tools in Excel

Hypothesis Testing: The Z-Test

A formal statistical test utilizing the standard normal distribution to evaluate hypotheses.

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

$\bar{X}$ = Sample Mean | μ_0 = Hypothesized Value | σ = Population Standard Deviation | n = Sample Size

- [✓] The population standard deviation (σ) must be known.
- [✓] OR the sample size n is sufficiently large ($n \geq 30$).
- [✓] The underlying data roughly follows a normal distribution.

Average testing tools in Excel

Hypothesis Testing: The t-Test

A hypothesis test deployed specifically when the true population standard deviation (σ) is unknown.

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

One-Sample t-test

Compares a single sample mean against a hypothesized constant (μ_0).

Independent t-test

Compares the means of two entirely independent, separate groups.

Paired t-test

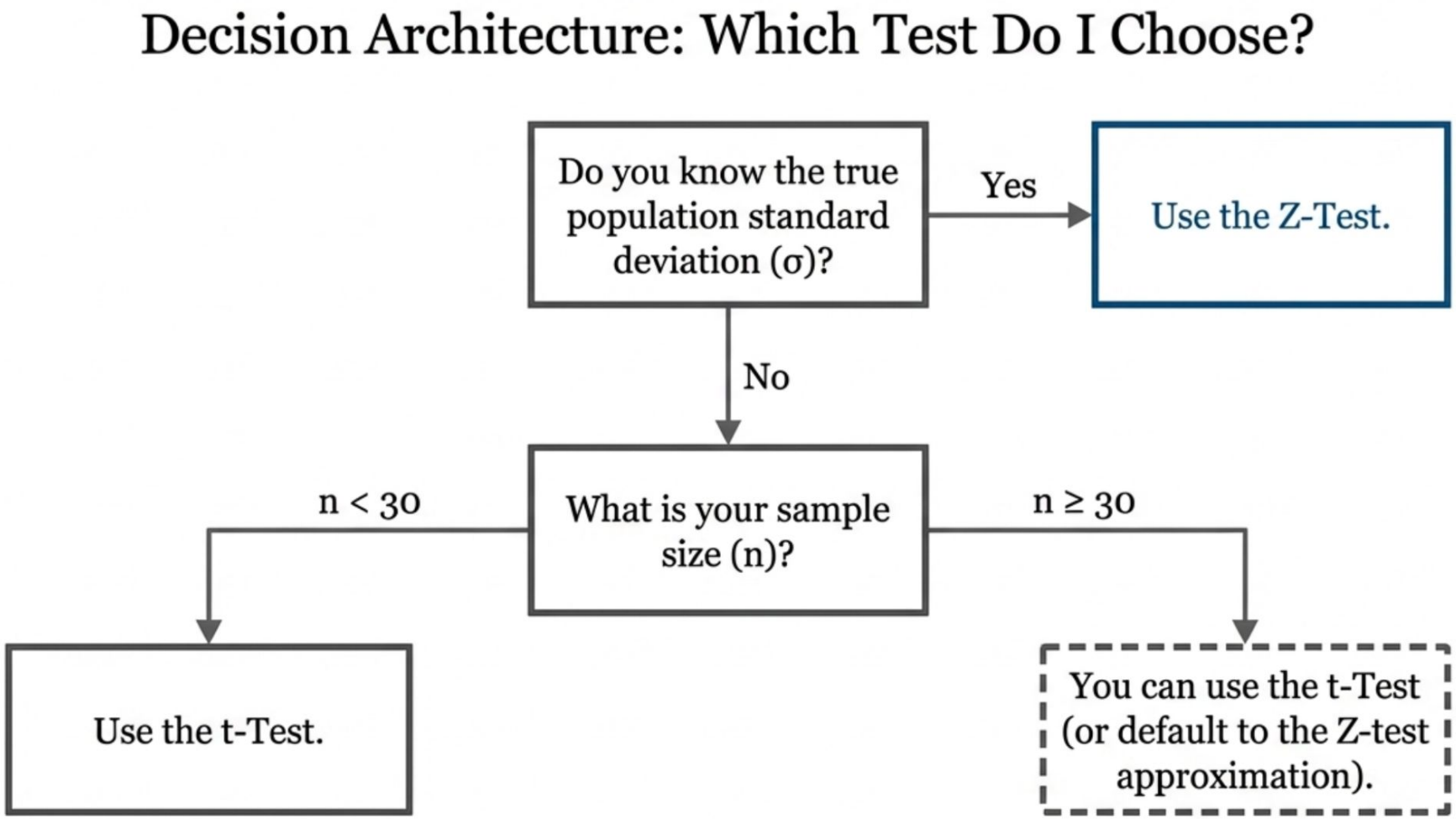
Compares two sets of paired observations (e.g., before/after measurements on the exact same subjects).

Average testing tools in Excel

Diagnostic Matrix: Z-Test vs. t-Test

Criteria	Z-Test	t-Test
Is population σ known?	Yes	**No**
Underlying Distribution	Standard Normal (Z)	Student's t
Sample Size Requirement	Large ($n \geq 30$)	Small or Any size
Distribution Tails	Thin	Thicker (Heavy Tails)
Operational Precision	High (because σ is known)	Highly practical (adapts to unknown σ)

Average testing tools in Excel



Average testing tools in Excel

"The t-test is essentially the uncertainty-aware version of the Z-test."

The Mathematical Reality

Because the math dictates that $t \approx Z$ when n is large, the t-test scales gracefully and automatically with the size of your data.

The Practical Reality

In real-world research, the true population standard deviation (σ) is **almost never known**.

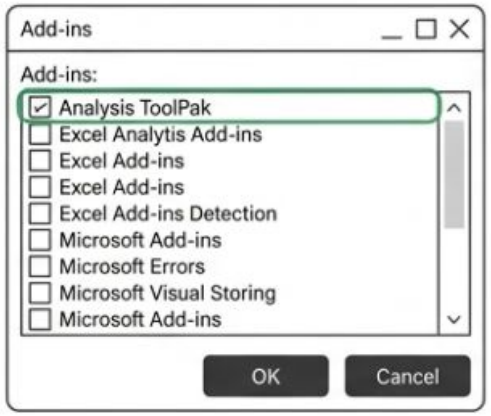
Consequently, the t-test is the overwhelmingly dominant choice in applied statistical research.

Average testing tools in Excel

Unlocking the Data Analysis ToolPak

Excel hides its most powerful reporting tools by default. Activating the ToolPak is the prerequisite for generating comprehensive statistical output reports.

File → Options → Add-ins → Excel Add-ins → Go → Analysis ToolPak



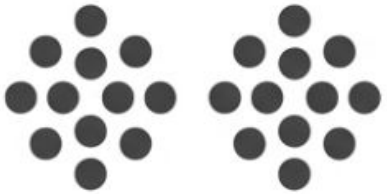
Average testing tools in Excel

A Comparative Overview of Excel's Three T-Tests

Paired Two Sample

Application: Comparing two observations on the exact same subject.

Example: Store sales before vs. after a remodel; app ratings before vs. after an update.



Equal Variances

Application: Comparing means of two independent groups assuming identical variance.

Example: Controlled A/B testing where both groups share identical demographic traits.

Unequal Variances (Welch)

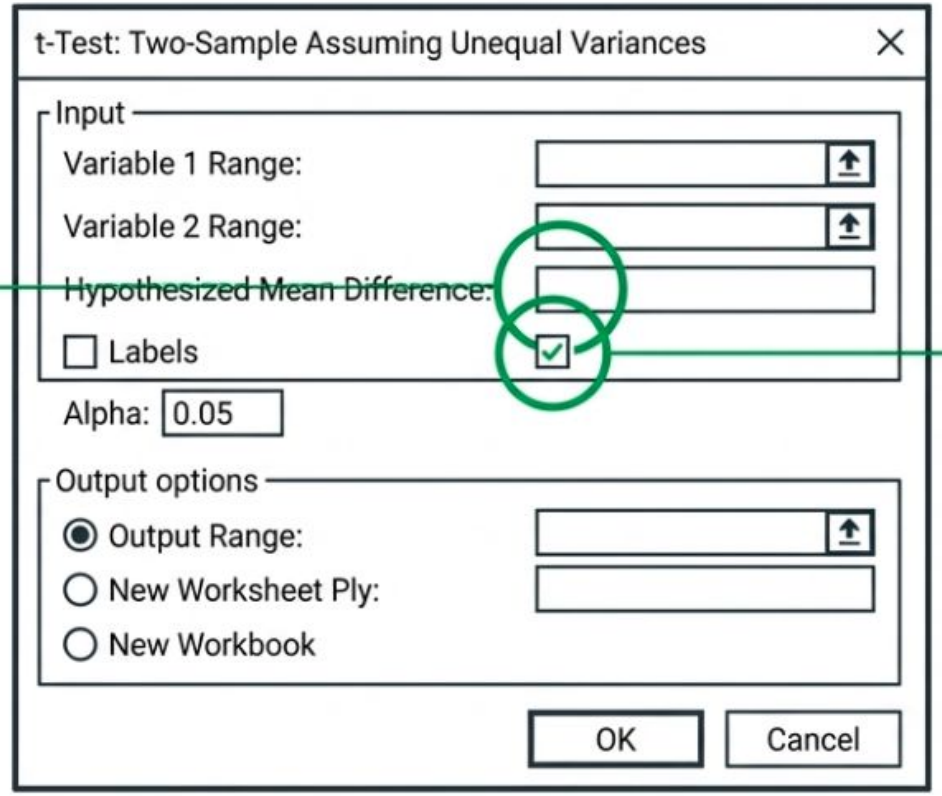
Application: Comparing independent groups with different sample sizes or volatilities.

The safest default choice when variance is uncertain.

Average testing tools in Excel

Configuring the ToolPak Dialog Box

Always enter 0. This mathematically represents the **Null Hypothesis (H0)** that there is exactly zero difference between the two means.



Always check the Labels box. Including column headers in your data range selection ensures the final generated report is instantly readable.

Average testing tools in Excel

Dynamic Dashboard Formulas

For live-updating reports where running the static ToolPak is impractical, utilize Excel's native array formulas.

```
=T.TEST(array1, array2, tails, type)
```

tails: 1 (one-tailed) or 2 (two-tailed).

type: 1 (Paired), 2 (Independent Equal Variance), 3 (Independent Unequal Variance).

Output: Returns the p-value directly.

```
=T.DIST.2T(t, df)
```

Output: Returns the two-tailed probability for the T-distribution when the t-stat and degrees of freedom are already known.

Average testing tools in Excel

The Two-Tailed Z-Test Formula Hack

Excel's native =Z.TEST function is inherently designed to return a one-tailed probability. To calculate a true two-tailed p-value for a Z-test, you must mathematically force the calculation using a nested MIN function.

```
=2 * MIN(Z.TEST(array, x, sigma), 1 - Z.TEST(array, x, sigma))
```

Multiplies the smallest tail by two, resulting in a symmetric two-tailed p-value.

Evaluates both tails and strictly grabs the smaller probability.

Forces Excel to calculate the complementary probability for the opposite tail.

Content

- Inferential statistics and hypothesis testing
- Average testing tools in Excel
- **Univariate linear regression**
- Output Summary

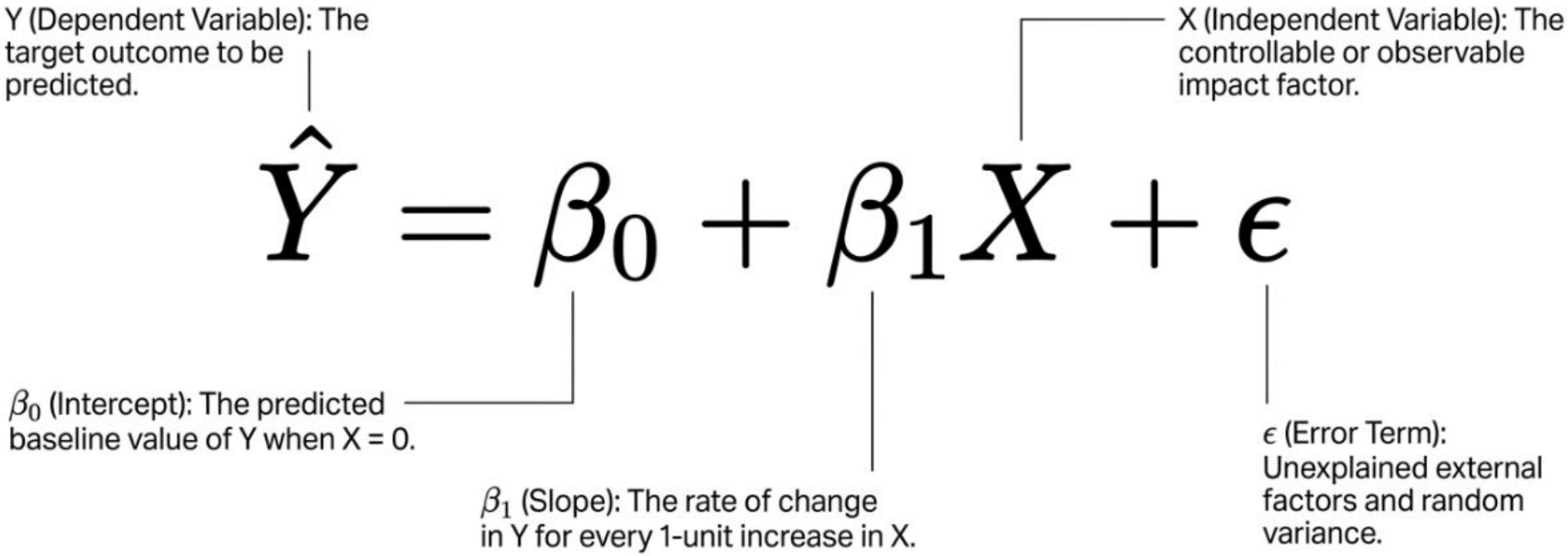
Univariate linear regression

The analytical leap from historical comparison to predictive modeling

The Past: Hypothesis Testing	The Present: Linear Regression
Core Question: "Is there a difference?"	**"If I change X, exactly how much will Y change?"**
Focus: Comparing historical states.	Focus: Quantifying cause-and-effect.
Limitation: Observes reality but does not quantify the mechanics of change.	Advantage: Serves as the foundation of Predictive Analytics in business.

Univariate linear regression

Deconstructing the regression equation into its functional components



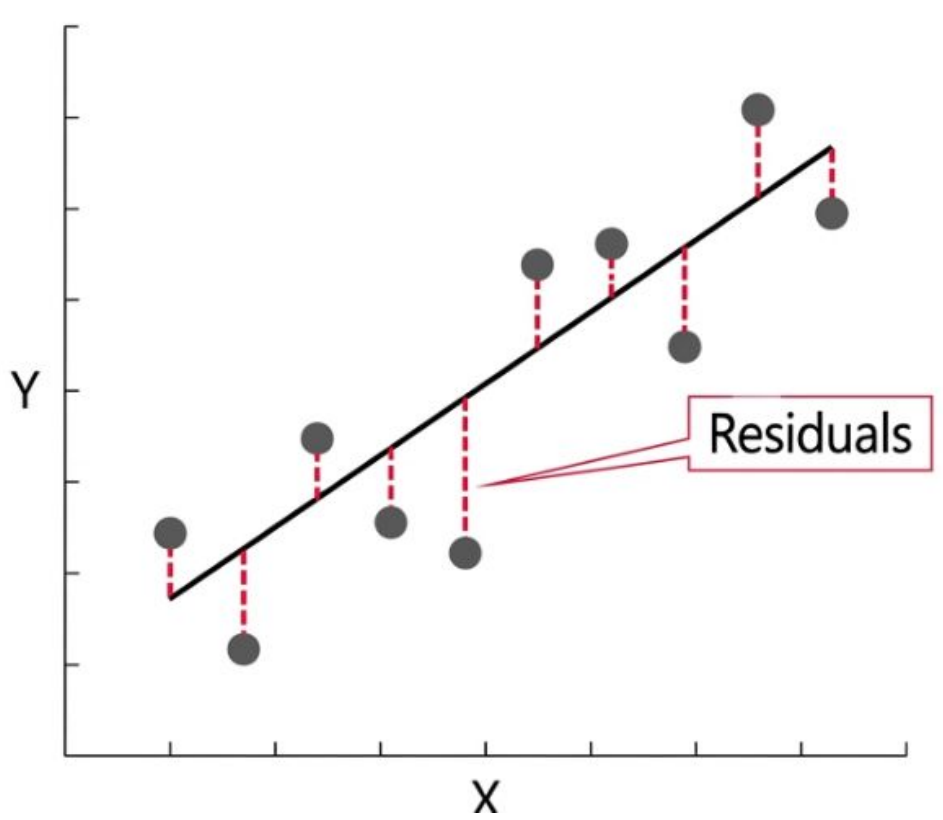
Univariate linear regression

Translating mathematical coefficients into actionable business metrics

β0 (Intercept)	→	**Fixed Costs** The baseline value that exists even when operational activity (X) is zero.
β1 (Slope)	→	**Variable Cost / Marginal Return** The exact incremental impact or cost incurred for every single unit of activity added.

Univariate linear regression

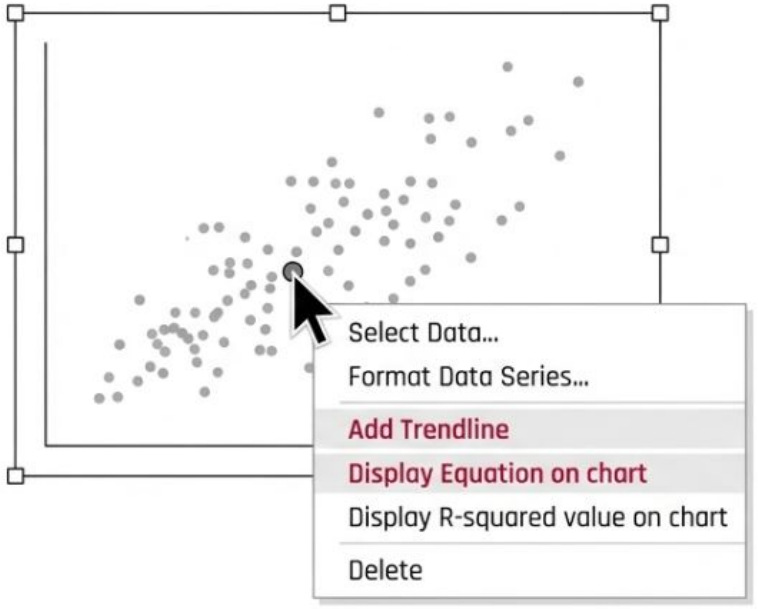
Ordinary Least Squares (OLS) minimizes the distance between prediction and reality



- **The Principle:** Excel mathematically tests thousands of potential lines through the data.
- **The Mechanism:** It calculates the vertical distance (Residual) between every actual data point and the predicted line.
- **The Goal:** The final model is the single line where the Sum of Squared Residuals is absolutely minimized, ensuring the tightest possible fit to reality.

Univariate linear regression

Establishing visual intuition before running matherïatical models



1. Never run complex regressions blind. Always visualize the relationship first.
2. Deploy a Scatter Plot to identify visual trends and potential anomalies.
3. Use Excel's quick chart tools to overlay a trendline and extract a preliminary equation to 'feel' the data's direction.

Univariate linear regression

The L.I.N.E. framework dictates the boundaries of model reliability

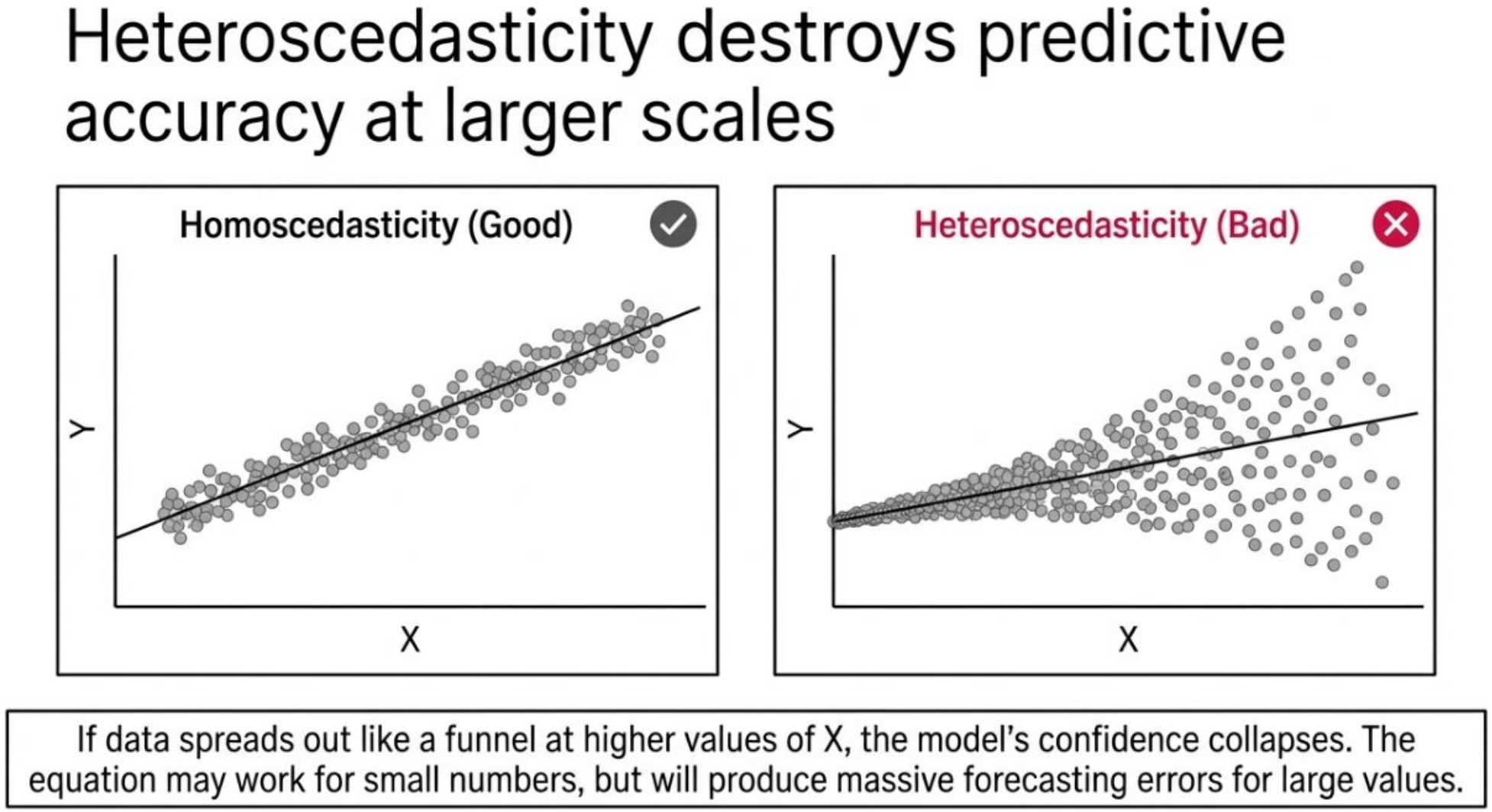
Linearity. The core relationship between X and Y must be a straight line, not a curve.

Independence. Error terms must not influence each other (critical for time-series data).

Normality. The residuals must follow a normal, bell-shaped distribution around the predicted line.

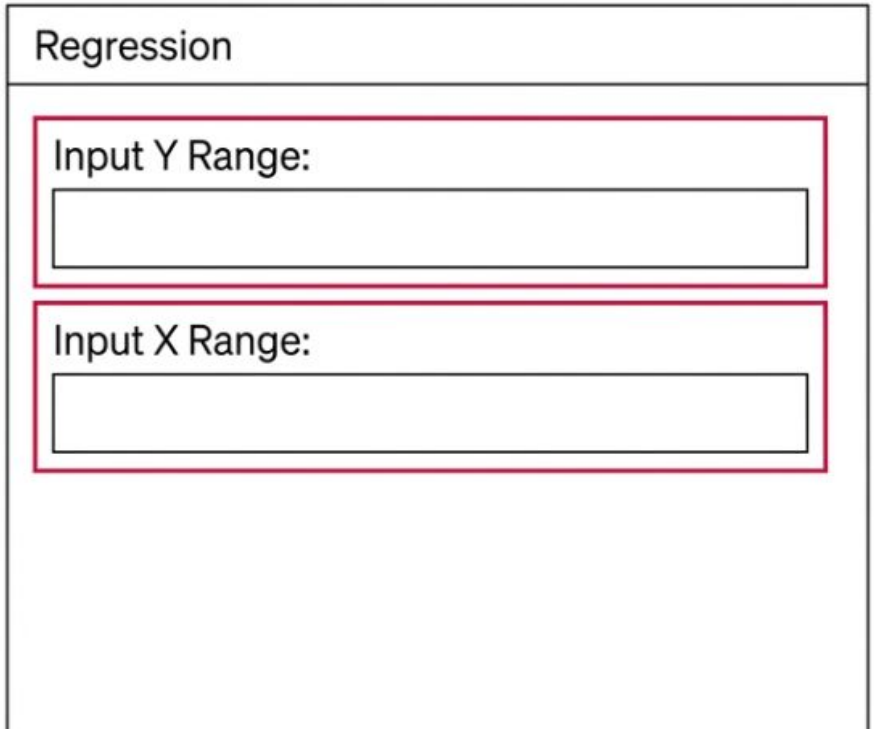
Equal Variance. The spread of data points around the line must remain uniform (Homoscedasticity).

Univariate linear regression



Univariate linear regression

Configuring the standard Excel environment for regression analysis

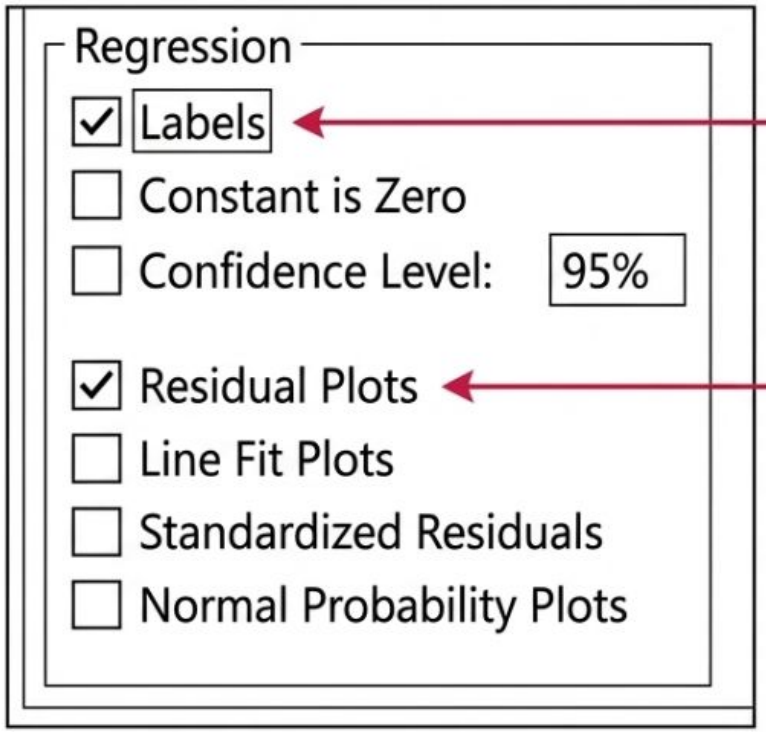


1. Step 1: Navigate to Data > Data Analysis > Regression.
2. Step 2 (Input Y): Select the column containing the Target outcome variable.
3. Step 3 (Input X): Select the column containing the Impact variable.

Crucial Note: Always include the column headers in your selection range.

Univariate linear regression

Mandating residual plots to ensure professional diagnostic rigor



Check 'Labels': Instructs Excel that the first row contains names, ensuring the final 'Summary Output' is readable and correctly assigned.

Check 'Residual Plots': Mandatory for professional analysis. This single checkbox generates the visual evidence required to verify the L.I.N.E. assumptions (specifically Equal Variance).

Step 5: Execute to generate the 'Summary Output' report.

Univariate linear regression

Deploying predictive models across core business functions

Operations Management

Predicting total production costs (Y) based on production volume (X).

Marketing

Forecasting revenue (Y) driven by digital ad spend budgets on Facebook/Google Ads (X).

Human Resources

Modeling specific salary bands (Y) determined strictly by years of professional experience (X).

Content

- Inferential statistics and hypothesis testing
- Average testing tools in Excel
- Univariate linear regression
- **Output Summary**

Output Summary

Reading the story behind the numbers makes you an analyst.

RAW DATA INPUT

20384, 99182, 10034, 76543, 00192, 44321, 88765, 11029, 33445,
66778, 55901, 22334, 77889, 00112, 44556, 88990, 11223, 33445,
66778, 55901, 22334, 77889, 00112, 00112, 44556, 88990, 11223,
33445, 66778, 55901, 22334, 77889, 00112, 44556, 88990, 11223,
33445, 66778, 55901, 11223, 33445, 33445, 66778, 55901, 22334,
33445, 66778, 55901, 22334, 77889, 00112, 44556, 44556, 88901,
22334, 77889, 00112, 46789, 00112, 44556, 88990, 11223, 33445,
66778, 55901, 22334, 22334, 77889, 00112, 44556, 88901, 77889,
00112, 44556, 88990, 11223, 33434, 77889, 00112, 44556, 88990,
11223, 33445, 66778, 55990, 00112, 44556, 44556, 88990, 11223,
33445, 66778, 55901, 77889, 00112, 00112, 44556, 88990, 11223,
77889, 00112, 00112, 44556, 88990, 11223, 33445, 66729, 33445,
33445, 66345, 55900, 00112, 44556, 88990, 77889, 00112, 44556,
22334, 77889, 00112, 44556, 88990, 11223, 33445, 66778, 55901,
88765, 11029, 33445, 66778, 55901, 77889, 00112, 44556, 88990,
00112, 33445, 22334, 77889, 00112, 44556, 88990, 11223, 33445,
66778, 55901, 77889, 00112, 29556, 91329, 33445, 66778, 55901,
22334, 77889, 00112, 44556, 00112, 44556, 88990, 11223, 33445,
00112, 44556, 22334, 77889, 00112, 44556, 88990, 11223, 33445,
33445, 66778, 88990, 11223, 33445, 66778, 55901, 22334, 77889,
33445, 66778, 55901, 22334, 77889, 00112, 44556, 88990, 11223

Needs interpretation

STRATEGIC ANALYSIS

This is where value is created: Understanding why.

☑ Running the regression makes you a data entrant.

Mechanical process, lacks insight.

☑ Decoding the output bridges the gap between statistical theory and strategic reality.

Insight informs strategy; data alone is just noise.

Output Summary

The Anatomy of a Regression Output

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.809224							
5	R Square	0.902324							
6	Adjusted R Square	0.909838							
7	Standard Error	0.865832							
8	Observations	80							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	277.856	1100.000	15.4507	0.0000			
13	Residual	122	396.476	0.4758					
14	Total	237	3529.310						
15									
		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
16	Intercept	-0.07207	0.04893	-1.3992	0.0000	-0.00644	1.09854	-0.00154	0.35290
19	Variable 1	0.05457	0.01615	-1.2785	0.0000	0.03128	0.36443	0.03286	0.35088
20	Variable 2	0.06178	0.05908	1.6646	0.0024	-0.00084	0.90368	0.09217	0.34292

Callout 1: Table 1 — Model Fit (Regression Statistics)

Callout 2: Table 2 — Global Significance (ANOVA)

Callout 3: Table 3 — The Equation (Coefficients)

Output Summary

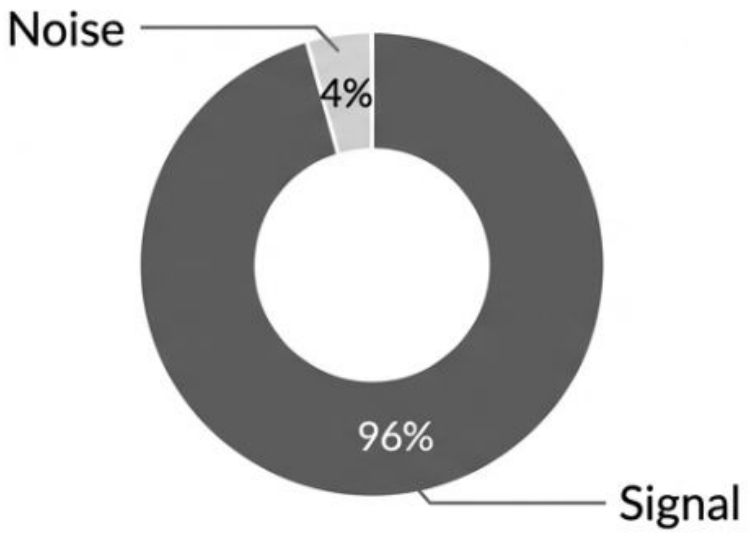
Three tables, three critical questions.

Table 1 (Fit)	How well does this model explain the data?	R Square
Table 2 (ANOVA)	Is the entire model actually meaningful?	Significance F
Table 3 (Coefficients)	How exactly do the variables interact?	Coefficients & P-values

Output Summary

R Square defines the explanatory power of your model.

Regression Statistics	
Pecidant	0.91
[illegible] with	0.7[illegible]
R Square	0.96
[illegible]	0.45[illegible]
HA	0
Value	0.57



An R^2 of 0.96 means 96% of the fluctuation in sales is directly driven by your variables (like advertising and price). Only 4% is attributed to random, unexplained factors.

Output Summary

The supporting metrics gauge precision and sample strength.

Regression Statistics	
Pecidant	0.91
R Square	0.96
Adjusted R Square	0.95
Multiple R	0.96
Standard Error	0.45
Observations	57

Multiple R

The Correlation Coefficient (-1 to 1). Measures the absolute strength of the linear relationship between X and Y.

Standard Error

The average deviation. Measures how far actual data points stray from your forecast line. **Smaller is always better.**

Observations

The total sample size fed into the analysis.

Output Summary

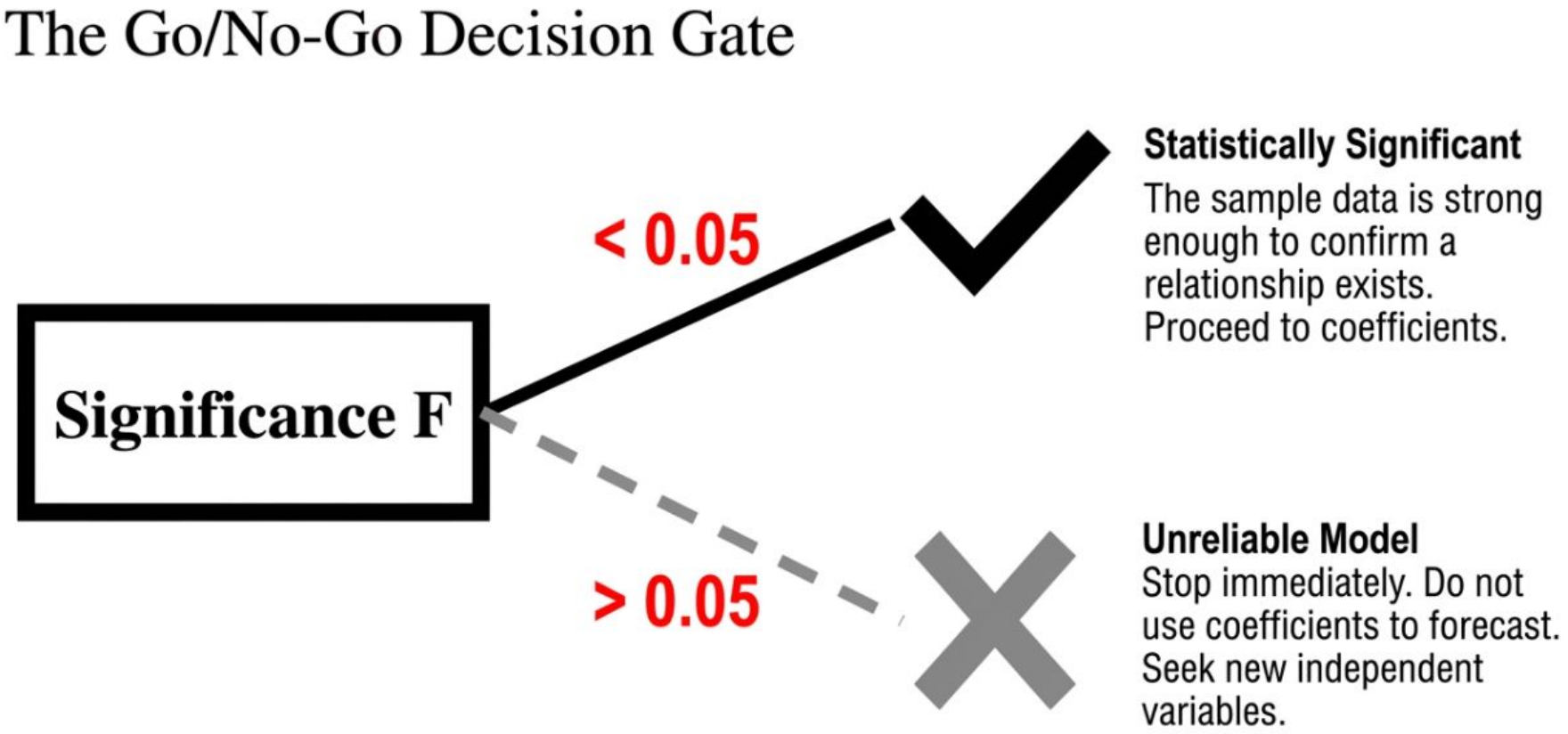
ANOVA tests if the entire model is statistically viable.

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.89E-24	1.36E-18	3.42004	1.34E-18
Residual	25	9.953.88	2.36E-22		
Total	187	13.82E-06			

The ANOVA table conducts a global hypothesis test (H_0). It assumes your model is useless and all regression coefficients are zero.

Significance F is the ultimate p-value that proves this assumption wrong.

Output Summary



Output Summary

Coefficients hold the mathematical soul of the model.

	Coefficients
Intercept	
X Variable	

Intercept

Math: Value of Y when X = 0.

Business: Fixed Costs or baseline revenue before any marketing impact.

X Variable

Math: Change in Y when X increases by 1 unit.

Business: Marginal Impact. The exact revenue generated per additional unit of investment.

Output Summary

Individual *P*-values dictate which variables to keep and which to cut.

Even if the global model passes ANOVA, each independent variable must prove its own worth.

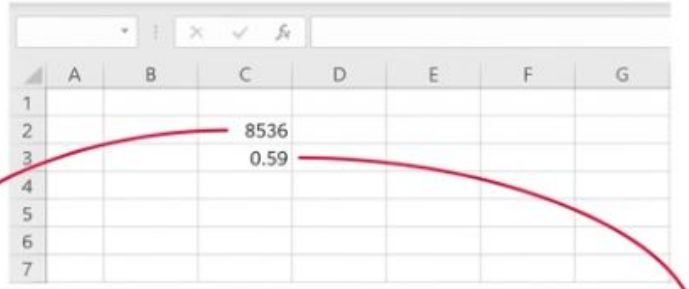
Table 3

	A	B	C	D
1		X	Y	P-value
2	Intercept	0.000	0.000	0.001
3	X Variable 1	-0.000	0.040	0.042
4	X Variable 2	-0.000	0.663	0.654
5				
...				

- **< 0.05**: The variable has a real, measurable impact. Keep it.
- **> 0.05**: The variable contributes no significant value. Consider removing it to make the model leaner and more accurate.

Output Summary

Building the predictive business equation.

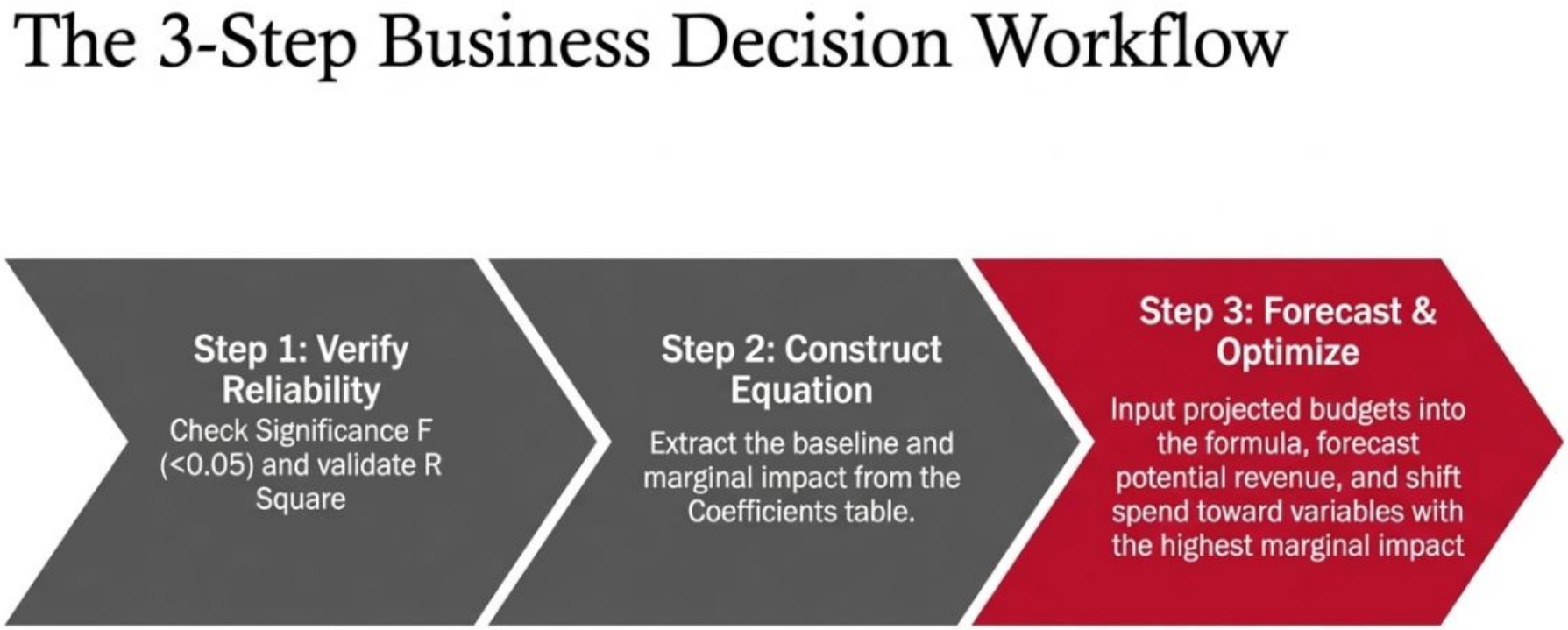


The Baseline (Intercept)

The Multiplier (Marginal Impact)

$$\text{Revenue} = 8536 + 0.59 \times \text{Advertising}$$

Output Summary



Thank you!