

Statistical Methods

Marvin Kweyu

2025-10-08

Table of contents

1	Preface	4
2	Introduction	6
2.1	How to use this guide	6
I	Distribution Tables	7
3	Discrete Distributions	8
4	Poisson Distribution	9
5	Continuous Distributions	10
6	Normal Distribution / Gaussian Distribution	11
7	Bernoulli Distribution	12
8	Binomial Distribution	13
9	Chi-Squared Distribution	14
10	F Distribution	15
11	Exponential Distribution	16
12	Logistic Distribution	17
II	Hypothesis Testing	18
13	Hypothesis testing	19
14	Analysis of Variance	20
15	One-Way ANOVA	21
16	Two-Way ANOVA	22

III Regression Analysis	23
17 Simple Linear Regression	24
17.1 Simple Linear Regression	24
18 Multiple Linear Regression	25
19 General Linear Models	26
20 Regression Summary	27
21 Sample Questions	28
22 Assessments	29
23 Recommended Reading	30
24 What Next	31

1 Preface

Preface

“There is a kind of knowledge that only comes from doing something the hard way, from first principles, with your own hands.”

~ Michael Mendy ([Linux from Scratch](#))

Statistical methods are rarely taught the way they are used. You encounter them in courses as isolated techniques - a t-test here, a regression there - each presented as though the hard part is remembering the formula. Then you sit in front of real data, with a real question, and realise the hard part was never the formula. It was knowing which one to reach for, why and what you are actually assuming when you do.

I started this guide because I needed one that did not exist. As a CS graduate daring to delve into the world of research, I had enough mathematical exposure to follow derivations but not enough statistical intuition to trust my own analysis. I could run the code, sure, but I could not always defend the output. Years of building production systems across agriculture, distributed infrastructure, and applied research had taught me a great deal about shipping software - and very little about the formal reasoning underneath the models I was increasingly relying on. That gap, between execution and understanding, is what this guide is an attempt to close.

This is not a passive document. I will reference books I am reading alongside this work. I will question assumptions I once treated as settled. I will break things deliberately to see where the edges are, and I will revise conclusions that no longer hold when examined more carefully. It will change as my understanding does - and as the field of statistical computing continues to move underneath all of us.

I cover:

- Probability foundations and distributional thinking
- Linear and generalised linear models
- Bayesian inference and prior specification
- Model diagnostics, selection, and validation
- Statistical computing in R

- Spatial and ecological applications

Most examples are drawn from work I have done or am currently doing: ecological niche modelling, geospatial data analysis, and applied research in climate and agriculture systems across Africa. The statistics here are not decorative.

Do not treat it as a reference manual. You will not find exhaustive API documentation or a formula sheet. What you will find is the reasoning behind the methods - why the assumptions matter, what breaks when they are violated, and what the output is actually telling you versus what it is tempting to believe it is telling you.

I dedicate this to the practitioners who *want to understand* what they are already doing, or want to do it with more honesty. Curiosity is assumed. Everything else, we build from here.

2 Introduction

There is a difference between doing statistics and understanding what you are doing. Most working data scientists live in that gap - running models, interpreting outputs, shipping results - without a clear account of the machinery underneath. This guide is an attempt to close that distance, not by simplifying the mathematics, but by grounding it in the context that makes it legible.

These pages cover statistical methods and statistical computing as they are actually used: imperfectly, iteratively, and usually under some form of deadline pressure.

2.1 How to use this guide

The guide assumes basic fluency in R, which remains the dominant language for statistical computing in research and academic contexts. If that assumption doesn't hold, [R for Data Science](#) is the right starting point - come back when variables and data frames feel natural.

Otherwise, read sequentially. The structure is deliberate: later sections assume earlier ones.

The two exceptions are:

1. You already know the preceding material and are targeting something specific.
2. A section explicitly signals it can stand alone - as in the case with Simple Linear Regression

The goal is not to replace a textbook. It's to be the thing you read alongside one - the part that explains why the textbook is saying what it's saying. Let's get started.

Part I

Distribution Tables

3 Discrete Distributions

4 Poisson Distribution

5 Continuous Distributions

6 Normal Distribution / Gaussian Distribution

7 Bernoulli Distribution

8 Binomial Distribution

9 Chi-Squared Distribution

10 F Distribution

11 Exponential Distribution

12 Logistic Distribution

Part II

Hypothesis Testing

13 Hypothesis testing

14 Analysis of Variance

15 One-Way ANOVA

16 Two-Way ANOVA

Part III

Regression Analysis

17 Simple Linear Regression

17.1 Simple Linear Regression

18 Multiple Linear Regression

19 General Linear Models

20 Regression Summary

21 Sample Questions

22 Assessments

23 Recommended Reading

Recommended Reading

24 What Next