# **Statistics in**

**Business, Finance,** 

**Management and** 

**Information Technology:** 

**A Layered Introduction** 

with **EXCEL** 

Draft Edition v0-26-1 Oct 2021

M. E. Kabay, PhD, CISSP-ISSMP Emeritus Professor of Computer Information Systems School of Cybersecurity, Data Science & Computing Norwich University

Free download:

http://www.mekabay.com/courses/academic/norwich/qm213/statistics\_text.pdf

## **PUBLICATION & COPYRIGHT STATEMENT**

This work-in-progress is copyright by its author, M. E. Kabay. The material is designed to support the author's QM213 course in the School of Business and Management at Norwich University. Version 0 is published only online as electronic files. The materials may be updated repeatedly during any semester; for example, the identifier

#### < statistics\_text\_v0-x-y.docx >

in the footer corresponds to the unpublished work (0) in the x<sup>th</sup> semester of use and the y<sup>th</sup> version created during that semester.

The work will eventually be printed and distributed commercially with all the exercises and question banks included electronically for other professors to use and modify easily.

#### **PERMISSIONS:**

- This text may be downloaded at no cost for the user. It may even be printed and distributed to students at the cost of production.
- It may not be posted online ANYWHERE other than the author's Website. It is practically impossible to keep versions up to date when they are scattered informally across the Internet.
- UNDER NO CIRCUMSTANCES MAY IT BE SOLD FOR PROFIT. The author would be intensely irritated to find someone ripping people off by selling what he provides free and would become quite nasty about it. Some creepy people print what's available free and then sell it on Amazon without permission. Any such abuse of innocent victims will result in extremely nasty lawsuits.

#### Dedication

To my beloved wife, Deborah Naomi Black light of my life;

and in gratitude to the professors whose devotion to clarity in teaching statistics set me on a life-long course of enthusiasm for the subject:

> Professor Hugh Tyson, formerly of McGill University,

> > and

Professors Robert R. Sokal z"l (deceased), and F. James Rohlf, both of State University of New York at Stony Brook, authors of the classic 1969 textbook, **Biometry, The Principles and Practice of Statistics in Biological Research**,

now in its Fourth Edition (2012).

# **Condensed Table of Contents**

Dedication	0-3
Condensed Table of Contents	0-4
Detailed Table of Contents	0-5
Preface	0-9
cknowledgements	0-13
Introduction	1-1
Accuracy, Precision, Sources of Data, Representing Data	2-1
Sorting, Backups and Enhanced Tables	3-1
Charts, Histograms, Errors in Graphing	4-1
Cumulative Frequency Distributions, Area under the Curve & Probability Basics	5-1
Descriptive Statistics	6-1
Sampling and Statistical Inference	7-1
Hypothesis Testing	8-1
Analyzing Relationships Among Variables	9-1
0 Analyzing Frequency Data	10-1
1 Introduction to Minitab 11-Error! Bookmark not	defined.
2 Multifactorial Analysis 12-Error! Bookmark not	defined.
3 Assumptions of Parametric Analysis	defined.
4 Exploratory Data Analysis	
5 Experimental Design Fundamentals	defined.
6 Miscellaneous Decision-Support Methods	defined.
7 Concluding Remarks	
8 Bibliography	

### Detailed Table of Contents<sup>1</sup>

Dedication	
Condensed Table of Contents	4
Detailed Table of Contents	5
Background	9
Why a "Layered" Introduction?	
Instant Tests	
The Importance of Homework	
Color vs Black-and-White	
Etymologies	
Question Authority	
Acknowledgements	

1	Introduction	1
1.1	About Applied Statistics	1
1.2	Computations in Applied Statistics	
1.3	Why EXCEL as Opposed to Other Statistical Packages?	4
1.4	Learning EXCEL	4
1.5	Long-Term Goals: A Personal Perspective	7
1.6	Using NUoodle	
1.7	SQ3R	9
1.8	Counting and Measuring	10
1.9	Variables	11
1.10	Tables	12
1.11	Choosing a Table Layout	
1.12	Transposing Rows and Columns	14
1.13	Types of Variables	
1.14	Qualitative / Categorical Data and Nominal Scales	16
1.15	Quantitative Data	
1.16	Discontinuous / Discrete Variables	17
1.17	Continuous Data	17
1.18	Interval Scales	
1.19	Ratio Scales	
1.20	Ordinal Scales: Ranks	19
1.21	Identifying the Type of Variable Really Matters to You	

2	Accuracy, Precision, Sources of Data, Representing Data	.1
2.1	Accuracy, Precision, and Being Correct	.1
2.2	Significant Figures	
2.3	Determining Suitable Precision for Statistics	
2.4	Sources of Real Statistical Data	
2.5	Representing Data	.9
2.6	Presenting Raw Data	

<sup>&</sup>lt;sup>1</sup> You may left-click any entry to jump to that page.

Sorting, Backups and Enhanced Tables	
Sorted Lists	
Simple Sorting in WORD	
Simple Sorting in EXCEL	5
Advanced Sorting in EXCEL	
Mistakes in Sorting	7
Making Backups of Your Work	
Enhancing the Presentation of Tables	
WORD Table Tools	
EXCEL Table Tools	
Copying EXCEL Tables into a WORD Document	
	Sorted Lists Simple Sorting in WORD Simple Sorting in EXCEL Advanced Sorting in EXCEL Mistakes in Sorting Making Backups of Your Work Enhancing the Presentation of Tables WORD Table Tools EXCEL Table Tools

4	Charts, Histograms, Errors in Graphing	
4.1	Horizontal Bar Charts vs Vertical Column Charts	
4.2	Pie Charts	
4.3	Clustered and Stacked Bar Charts and Column Charts	6
4.4	Creating Charts in WORD	
4.5	Managing Figure & Table Numbers in WORD	9
4.6	Editing Graphics in EXCEL	
4.7	Frequency Distributions	
4.8	Histograms	
4.9	Creating Frequency Distributions and Histograms in EXCEL	
4.10		
4.11	Problems with Disparate Quantities	
4.12		
4.13	Truncating the Ordinate	
4.14		

#### 5 Cumulative Frequency Distributions, Area under the Curve & Probability Basics ...... 1

5.1	Relative Frequencies, Cumulative Frequencies, and Ogives	1
5.2	Area under the Curve	3
5.3	Basic Concepts of Probability Calculations	5
5.4	The Uniform Probability Distribution	9
5.5	The Normal Probability Distribution	11
5.6	Area under the Curve for Any Normal Distribution	13
5.7	Area Under the Curve for the Standard Normal Distribution	14
5.8	Using EXCEL Functions for Areas Under Other Probability Distribution Curves	15
5.9	Chi-Square Distribution	16
5.10	F Distribution	18
5.11	Student's-t Distribution	19

6	Descriptive Statistics	1
---	------------------------	---

	-	
6.1	Summarizing Groups of Data using EXCEL Descriptive Statistics	1
6.2	Computing Descriptive Statistics using Functions in EXCEL	3
6.3	Statistics of Location	4
6.4	Arithmetic Mean ("Average")	4

6.5	Calculating an Arithmetic Mean from a Frequency Distribution	
6.6	Effect of Outliers on Arithmetic Mean	6
6.7	Median	
6.8	Quantiles	9
6.9	EXCEL 2010 .INC and .EXC Functions	
6.10	Quartiles in EXCEL	11
6.11	QUARTILE.EXC vs QUARTILE.INC	12
6.12	Box Plots	13
6.13	Percentiles in EXCEL	14
6.14	Rank Functions in EXCEL	
6.15	Mode(s)	17
6.16	Statistics of Dispersion	19
6.17	Range	19
6.18	Variance: $\sigma^2$ and $s^2$	
6.19	Standard Deviation: $\sigma$ and s	21
6.20	Skewness	
6.21	Kurtosis	23

7	Sampling and Statistical Inference	1
7.1	Populations and Samples	1
7.2	Sample Statistics and Parameters	
7.3	Greek Letters for Parametric Statistics	
7.4	Random Sampling from a Population	4
7.5	Selecting Random Values for an Unbiased Sample	7
7.6	More about Probability and Randomness	
7.7	Random Number Generators	
7.8	Probabilities in Tossing Coins	10
7.9	Probabilities in Statistical Inference	10
7.10	The Central Limit Theorem in Practice	11
7.11	The Expected Value	
7.12	More About the Normal Distribution	
7.13	Statistical Inference: Interval Estimation	
7.14	Population Mean Estimated Using Parametric Standard Deviation	
7.15	Estimating Parametric Mean Using the Sample Standard Deviation	20
7.16	Degrees of Freedom Vary in Statistical Applications	
7.17	Notation for Critical Values	
7.18	Two-Tailed Distributions	
7.19	EXCEL CONFIDENCE.T Function	
7.20	Beware the Definition of a in Inverse Probability Functions	
7.21	Interval Estimate for Any Normally Distributed Statistic	
7.22	Population Proportion Based on Sample Proportion	26
7.23	Conditional Formatting	
7.24	Confidence Limits for Population Variance and Population Standard Deviation Based on Samp	le
Varia	bilitybility	29

8	Hypothesis Testing	. 1
8.1	Introduction	.1
8.2		

8.3	Levels of Statistical Significance and Type I Error: Rejecting the Null Hypothesis When it is	
Actual	ly True	5
8.4	Type II Error: Accepting the Null Hypothesis when it is Actually False	7
8.5	Testing a Sample Variance Against a Parametric Value	7
8.6	Are the Means of These Two Populations the Same?	9
8.7	ANOVA for Comparing Means of Two Samples	9
8.8	The Model for Single-Factor ANOVA	
8.9	Testing for the Equality of Two Means in an Investigation of Possible Dishonesty	15
8.10	T-Tests in Data Analysis	16
8.11	Critical Values	18
8.12	ANOVA: Single Factor vs T-test for Equality of Means	19
8.13	Testing for Equality of Means Given Parametric Mean and Parametric Standard Deviation v	
Sample	e Mean	20
8.14	Computing a t-test for Equality of Means without Raw Data	21
8.15	The T.TEST Function	

9	Analyzing Relationships Among Variables	. 1
9.1	Introduction to Analyzing Relations	. 1
9.2	Cross-Tabulations (Contingency Tables)	
9.3	Filtering Data for Temporary Views	
9.4	Charts for Contingency Tables	
9.5	Scatterplots and the Intuitive Grasp of Relationships	
9.6	Pearson Product-Moment Correlation Coefficient, r	
9.7	Computing the Correlation Coefficient Using EXCEL	
9.8	Testing the Significance of the Correlation Coefficient	
9.9	Coefficient of Determination, $r^2$	
9.10	Linear Regression in EXCEL	
9.11	ANOVA with Linear Regression	
9.12	Predicted Values in Linear Regression & Confidence Limits	
10	Analyzing Frequency Data	. 1
10.1	Computing Expected Frequencies for Statistical Distributions	. 1
10.2	The Chi-Square Goodness-of-Fit Test	
10.3	The EXCEL =CHISQ.TEST Function	
10.4	Two-Way Tests of Independence Using Chi-Square	5
[Chapter	r headings for future expansion of text are omitted here]	
18	Bibliography	. 1

## Preface

I hope that students will enjoy their introduction to applied statistics. To that end, the course and this text are designed with learning in mind. The unusual layered approach is the expression of my almost 50 years of teaching (I started in 1963): instead of drowning students in increasingly bewildering detail for each topic, I want to start with WHY they should learn the material and then show them comprehensible, manageable chunks of practical, useful concepts and techniques. With some practical knowledge and techniques mastered, they can then come back to what they have already started to learn to fill in additional details using the their foundation for easier comprehension of subtleties.

# Background

Students and teachers may be interested in knowing how a professor of information systems and information assurance also came to be a fanatic about applied statistics. If not, just skip to the next section!

In 1969, when I was a student in the Department of Biological Sciences at McGill University in Montreal, Canada, Dr Hugh Tyson taught an introduction to biostatistics using the first edition of Robert R. Sokal and F. James Rohlf's *Biometry* text. The course thrilled me. I use the verb deliberately: it struck a deep chord of delight that combined my love of biology with my life-long enthusiasm for mathematics. I had completed high school math by the age of nine and taught seniors in my high school matriculation math by the age of 13; they used to call me "Slide Rule" because I carried one on my belt. My master's thesis was a statistical analysis of extensive data collected by my research director, Dr Daphne Trasler, a renowned teratologist in the Human Genetics Sector at McGill University, about the developmental variability of inbred mouse strains and of their hybrids.

At Dartmouth College, I was admitted to the doctoral program because I helped a world-famous invertebrate zoologist apply appropriate analytical methods to frequency data, resulting in accurate estimates of the probability of the null hypotheses. Because of my knowledge of statistics, I was granted a full four-year waiver of tuition and was paid generously through the program as a Teaching Assistant and then Research Assistant, finally being given permission to teach an informal graduate seminar on applied statistics to my fellow graduate students. I served as a statistical consultant to several professors for their experimental design and data analysis. My PhD oral field exam was in applied statistics and invertebrate zoology.

After my thesis was accepted in August 1976, I was hired by the Canadian International Development Agency in October 1976 to teach three levels of applied statistics in the Faculty of Social and Economic Sciences at the National University of Rwanda (in French, my native language); I was delighted to be asked by the Faculty of Agronomy also to teach a course on field experiment design and by the Faculty of Science to teach FORTRAN programming. On my return from Africa in 1978, I then taught biostatistics and some biology courses at the University of Moncton (also in French) for a year.

Although I began programming in 1965 as a kid, my first formal job in the computer science field was as a programming statistician: I was hired in 1979 (thanks to a recommendation from a friend I had known in graduate school) to create, define and parse the statistical syntax for INPROSYS, the compiler for a new fourth-generation language and relational-database system, and to write the code generator for that syntax.

All through my career since then, I have served as a statistical consultant to colleagues and especially for my wife, Dr Deborah N. Black, MDCM, FRCP(C), FANPA, who has graciously named me as coauthor for some of her papers. At the National Computer Security Association, where I served as Director of Education between 1991 and 1999, I was also responsible for ensuring the survey design and statistical rigor of several of our annual virus-prevalence surveys.

At Norwich University, I was delighted to substitute-teach the QM370 *Quantitative Methods for Marketing & Finance* course in Spring 2002 for a professor on sabbatical and then to be offered the chance to teach QM213 *Business and Economic Statistics I* in Spring 2010. With the support of the Directors of the School of Business and Management, I hope to continue teaching QM213 until I retire on June 30, 2022! **Students**: knowing statistics in addition to your main focus of study well REALLY HELPS YOU in your career! Take this course seriously!

# Why a "Layered" Introduction?

In my experience of teaching statistics, I have found that textbooks are often designed as if they were reference books. They dive into depth on every topic in turn, bewildering, exhausting, and dispiriting students, who get lost in detail without grasping why the material should matter to them in their academic or professional work.

Teaching style should avoid overload and should motivate interest, giving students the opportunity to form a network of firm associations among concepts and new vocabulary before plunging into sophisticated detail.

Nothing except conservatism and tradition – or, in the words of Monty Python's *Architect Skit*, "blinkered, Philistine pig-ignorance"<sup>2</sup> stops us from introducing interesting and valuable concepts and techniques and then returning for deeper analysis and knowledge once students have begun building their own conceptual and experiential framework.

In addition, *forward references* to subjects that will be explored later in the course are valuable to students as a basis for forming increasingly complex neural networks that facilitate absorption of details later in their studies. For example, getting students used to the names and applications of analysis of variance, regression, non-parametric statistics, and other topics helps them when they plunge into the details, computations and interpretations of these methods. Instead of having to assimilate everything at once – the existence of the method, its name, its application, its computation, and its interpretation – the students have an Ahal experience as they reach the section about something they've heard about several times before.

Students in Norwich University's QM213 Business & Economic Statistics preferring to use a paper copy instead of or in addition to the electronic version is welcome to ask for one and I'll print it at no cost to them thanks to the kindness of the University administration.

#### **Instant Tests**

Following up on a suggestion from QM213 student Dejan Dejan, who took the course in Fall 2010, I have inserted boxes with a few review questions throughout the early part of the text. Most of the questions are conceptual; some suggest little real-world research tasks for posting in the NUoodle classroom discussions; some have computational exercises. There are no answers posted in the book, requiring students to compare notes with each other – a Good Thing. If students are stumped they should review the text, discuss the questions with fellow-students, and ask the professor for help. And students should note that suggestions for improvement are always welcome!

# The Importance of Homework

Reading about methods is too abstract to grip students emotionally or to solidify the engrams (memory traces) that underlie learning. Practical application of these techniques using interesting cases stimulates the imagination and builds neural patterns that make it easier to learn new statistical concepts and techniques.

Combining practice with repeated exposure to concepts through a layered approach to teaching helps students convert short-term memory into long-term knowledge. In my statistics courses over more than 30 years of experience, I have always assigned half or more of the final grade to homework.

<sup>&</sup>lt;sup>2</sup> < <u>http://www.youtube.com/watch?v=e2PyeXRwhCE</u> >

Remember: reading about statistics must surely be the most passionately interesting and absorbing activity in your life right now O but the only way to be good at statistics – and to pass this course – is to *do the homework*.

A practical problem arose: with 70 students in a statistics course and each one doing half-a-dozen problems per week, how can one professor grade the homework?

- Trying to grade all the results myself proved impossible it took longer than a week to grade a single week's work.
- In the next statistics class, I tried having the students grade their own homework in class and ended up spending one day out of three just doing the grading!
- In the next course sessions, I tried having the students do the homework and then answer questions about the specific values in particular aspects of their work. However, another problem developed in the Fall 2011 and Spring 2012 sessions of QM213, when I succumbed to an excess of sympathy for the piteous pleas of the students and allowed them to try randomized quizzes and homework assignments up to three times, taking the best of the results for each student. Unfortunately, some students were gaming the system by recording the correct answers supplied by the NUoodle system and using the lists of correct inputs to improve their scores without actually studying or doing homework at all. Students can thank those who chose to cheat instead of to study for the reduction of homework and quiz assignments to a single try.
- In addition, thanks to a student in one class who claimed to have completed all the homework despite evidence from the NUoodle log files that he never even logged into the NUoodle group between mid-February and early May, I am requiring that a selection of homework files be uploaded upon completion to provide an augmented audit trail.

Finally, several students and I quickly realized that in this course, falling behind can have disastrous consequences: what should be an easy extension of concepts and techniques mastered earlier becomes a morass of increasingly incomprehensible gibberish.

The solution starting in QM213 for the Fall of 2012 is to assign the readings and homework at the start of each week using the NUoodle online learning system to test each student on the material by the end of the second Sunday after the start of the week.

To help students who are having trouble grasping the concepts or mastering the techniques, I am providing compensatory (replacement) homework and exams that allow students to improve their scores by replacing the earlier bad scores by the later better (we hope) scores to demonstrate improvements. These replacement homework assignments and replacement quizzes are opened a few weeks after the initial assignments so students having a hard time can ask for help.

As always in my courses, I support the work of the Academic Achievement Center (AAC) at our University and provide "a" versions of all quizzes and grading with twice the usual time allowance. One learning-disabled student actually came 1<sup>st</sup> in his class. I don't care whether you learn fast or slow: I just want to support your learning, no matter what!

Finally, throughout this book and the corresponding course, I introduce topics only because **they make a difference in practical applications of statistics**. As I always say, REALITY TRUMPS THEORY. There is no content presented "because it's good for you" or "because I had to study this decades ago and therefore I'm going to force you to learn it too even though it makes no difference to anyone today." You will note that I do *not* ask you to memorize algebraic formulas; when formulas are presented they are for explanatory purposes. This is not a math course, and you will never be asked to memorize derivations of formulas. And you absolutely will not look up statistical critical values in outdated tables: all the statistical functions needed for a basic level of applied statistics are available in statistical packages and in EXCEL 2007, 2010, 2013 and 2016 in particular. Mac users must install their own version of the statistical analysis pack or simply use the Windows version of EXCEL available on any University computer.

# Color vs Black-and-White

This textbook is designed for people who can see color, but should also be accessible to color-blind readers to whom the colors will appear as different shades of gray. The full-color version is available as a PDF on the course Website. When the textbook passes beyond the v0.x stage and is ready for a first formal printing, it will be available as book printed in color. Later editions will be supplied with sound recordings of the text to help visually impaired students and for general application in review (e.g., listening to chapters while travelling – or as a perfect way of being put to sleep).

# Etymologies

Throughout the text, you will occasionally encounter footnotes that explain the origins (etymology) of technical terms. You are not expected to memorize any of these! They're added simply as a tool to encourage students to learn and remember Greek and Latin roots that are often used in technical terminology. With experience, you may be able to at least guess at the meaning of phrases like "stygian obscurity" and "Sisyphean futility."

# **Question Authority**

**If you don't understand something, ASK!** In all the years that I have been teaching (since 1963, when I tutored seniors in my high school who were failing their matriculation examinations), I have *never* criticized, sneered at, embarrassed or humiliated a student for wanting to understand something! And there's an excellent chance that someone else in class has exactly the same question but hasn't asked it yet. Think of all the time you can save for yourself and others simply by being unembarrassed and frank.

If you doubt my assertion, please speak with students who have completed other courses with me; I am confident that they will confirm that I am not a \*\*\*\*ing \*%\*#%\$ who abuses students!

In class, I often ask the class for snap judgements and often precede them with the comment that I don't care if you are right or wrong – I'm just trying to keep your brains active. If you're right, great! If you're not, you're learning without pain.

Don't *ever* hesitate to ask a question in class, after class, by visiting me in my office, by Skype or by phone. I work for *you* and get enormous pleasure from helping people *get it*. If a professor (me too) ever says something you don't understand, be sure to clarify the issue at an appropriate time. Don't give up, ever.

In addition, students will quickly discover that I respond positively to constructive suggestions for improving the textbook, the exercises, the structure of the teaching system, homework, exams and examples. I keep a record of corrections and suggestions for improvement in the CONTINUOUS PROCESS

IMPROVEMENT discussion group in NUoodle and grant extra points for such contributions. Finding and fixing errors are not embarrassing: I wholeheartedly support continuous process improvement and regard huffy resistance to corrections or positive suggestions for improvement as an indication of mental rigidity, outright stupidity, or neurotic insecurity.

Exceptionally good contributions may even get you mentioned in the acknowledgements, as you can see for yourself.

#### (38)

### Acknowledgements

As indicated in the Dedication page, it is impossible fully to describe the degree of support poured out upon me by my wife, Dr Deborah Black during this years-long project. She has not only tolerated the hours of absence as I slaved over the text but has enthusiastically greeted the latest cry of, "Another chapter done!" – for summer after summer.

I am eternally grateful to Dr Hugh Tyson and Dr Daphne Trasler, formerly of McGill University for their teaching and encouragement to pursue statistics professionally. Dr John J. Gilbert of Dartmouth College was an important force in my scholarly evolution and gave me tremendous support in my drive to become a statistician.

Former Dean Dr Frank Vanecek, statistician and economist Dr Mehdi Mohaghegh, and econometrician Dr Najiba Benabess (now Dean of the Tabor School of Business at Milliken University) of the School of Business and Management at Norwich University have showered me with enthusiastic encouragement and support of this project from the very beginning. Prof David Blythe has continued to allow me to teach the course every year – and recently, every semester!

The 13 students in the Fall 2010 semester of the experimental session of QM213 at Norwich University in the School of Business and Management who used the first version of this text contributed many practical suggestions for improvement – either explicitly or simply through the difficulties they experienced with different aspects of the subject matter. Student Dejan Dejan in particular had many helpful and articulate suggestions for improvement of the text and of the course and I am particularly grateful to him. Many other students in the sessions since then have also been helpful and cooperative in locating errors and suggesting corrections and other improvements.

Finally, I must mention with gratitude Pop (Percy Black z"l) and Mom (Virginia Black), distinguished professors and professional writers themselves, whose constant encouragement was a real Wonderbra experience (i.e., "uplifting," as Pop used to say).

Naturally, all errors are my own fault and, in the tradition of Monty Python, I grovel in the reader's general direction.<sup>3</sup> Please send corrections and suggestions for improvement to me through the NUoodle classroom' *Continuous Process Improvement* section if you are one of my students or by external email if you are reading, studying, or using the text at another institution or on your own.

Much

M. E. Kabay Northfield, Vermont April 2019 \* \* \*

NORWICH EMAIL for Norwich students, staff and faculty: < <u>mkabay@norwich.edu</u> > GMAIL for everyone else: < <u>mkabay@gmail.com</u> >

<sup>&</sup>lt;sup>3</sup> Based on Monty Python's Holy Grail, Scene 8. French Guard: "I fart in your general direction."

<sup>&</sup>lt; http://arago4.tnw.utwente.nl/stonedead/movies/holy-grail/main.html >