

Statistics in Business, Finance, Management and Information Technology: A Layered Introduction with EXCEL, WORD & Minitab

**Draft Edition v0-25-2
April 2019**

**M. E. Kabay, PhD, CISSP-ISSMP
Professor of Computer Information Systems
School of Business and Management
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in the footer corresponds to the unpublished work (0) in the xth semester of use and the yth version created during that semester.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 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
Acknowledgements.....	0-13
1 Introduction.....	1-1
2 Accuracy, Precision, Sources of Data, Representing Data	2-1
3 Sorting, Backups and Enhanced Tables.....	3-1
4 Charts, Histograms, Errors in Graphing.....	4-1
5 Cumulative Frequency Distributions, Area under the Curve & Probability Basics	5-1
6 Descriptive Statistics	6-1
7 Sampling and Statistical Inference.....	7-1
8 Hypothesis Testing	8-1
9 Analyzing Relationships Among Variables.....	9-1
10 Analyzing Frequency Data.....	10-1
11 <i>Introduction to Minitab</i>	11-1
12 <i>Multifactorial Analysis</i>	12-1
13 <i>Assumptions of Parametric Analysis</i>	13-1
14 <i>Exploratory Data Analysis</i>	14-1
15 <i>Experimental Design Fundamentals</i>	15-1
16 <i>Miscellaneous Decision-Support Methods</i>	16-1
17 <i>Concluding Remarks</i>	17-1
18 Bibliography	18-1

Detailed Table of Contents¹

Dedication.....	3
Condensed Table of Contents.....	4
Detailed Table of Contents	5
Background.....	9
Why a “Layered” Introduction?.....	10
Instant Tests.....	10
The Importance of Homework.....	10
Color vs Black-and-White.....	11
Etymologies.....	12
Question Authority.....	12
Acknowledgements.....	13
 1 Introduction.....	1
1.1 About Applied Statistics	1
1.2 Computations in Applied Statistics.....	4
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.....	8
1.7 SQ3R	9
1.8 Counting and Measuring	10
1.9 Variables.....	11
1.10 Tables.....	12
1.11 Choosing a Table Layout.....	13
1.12 Transposing Rows and Columns	14
1.13 Types of Variables	16
1.14 Qualitative / Categorical Data and Nominal Scales	16
1.15 Quantitative Data	16
1.16 Discontinuous / Discrete Variables	17
1.17 Continuous Data.....	17
1.18 Interval Scales.....	18
1.19 Ratio Scales	18
1.20 Ordinal Scales: Ranks.....	19
1.21 Identifying the Type of Variable Really Matters to <i>You</i>	20
 2 Accuracy, Precision, Sources of Data, Representing Data	1
2.1 Accuracy, Precision, and Being Correct	1
2.2 Significant Figures	2
2.3 Determining Suitable Precision for Statistics.....	3
2.4 Sources of Real Statistical Data.....	6
2.5 Representing Data	9
2.6 Presenting Raw Data.....	9

¹ You may left-click any entry to jump to that page.

3	Sorting, Backups and Enhanced Tables	1
3.1	Sorted Lists	1
3.2	Simple Sorting in WORD	1
3.3	Simple Sorting in EXCEL.....	5
3.4	Advanced Sorting in EXCEL.....	6
3.5	Mistakes in Sorting	7
3.6	Making Backups of Your Work	8
3.7	Enhancing the Presentation of Tables	10
3.8	WORD Table Tools.....	10
3.9	EXCEL Table Tools.....	15
3.10	Copying EXCEL Tables into a WORD Document.....	16
4	Charts, Histograms, Errors in Graphing.....	1
4.1	Horizontal Bar Charts vs Vertical Column Charts.....	1
4.2	Pie Charts.....	3
4.3	Clustered and Stacked Bar Charts and Column Charts.....	6
4.4	Creating Charts in WORD.....	8
4.5	Managing Figure & Table Numbers in WORD	9
4.6	Editing Graphics in EXCEL	10
4.7	Frequency Distributions	11
4.8	Histograms.....	12
4.9	Creating Frequency Distributions and Histograms in EXCEL	13
4.10	Choosing a Reasonable Number of X-axis Values	17
4.11	Problems with Disparate Quantities.....	18
4.12	Logarithmic Scale on the Ordinate.....	20
4.13	Truncating the Ordinate	23
4.14	Selecting Non-Random Sections of a Data Series.....	25
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	5
6.6	Effect of Outliers on Arithmetic Mean	6
6.7	Median.....	8
6.8	Quantiles	9
6.9	EXCEL 2010 .INC and .EXC Functions.....	10
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.....	14
6.15	Mode(s)	17
6.16	Statistics of Dispersion	19
6.17	Range	19
6.18	Variance: σ^2 and s^2	20
6.19	Standard Deviation: σ and s	21
6.20	Skewness	22
6.21	Kurtosis.....	23
7	Sampling and Statistical Inference.....	1
7.1	Populations and Samples.....	1
7.2	Sample Statistics and Parameters	2
7.3	Greek Letters for Parametric Statistics	3
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	9
7.7	Random Number Generators.....	10
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.....	13
7.12	More About the Normal Distribution	13
7.13	Statistical Inference: Interval Estimation.....	16
7.14	Population Mean Estimated Using Parametric Standard Deviation	17
7.15	Estimating Parametric Mean Using the Sample Standard Deviation.....	20
7.16	Degrees of Freedom Vary in Statistical Applications	21
7.17	Notation for Critical Values.....	21
7.18	Two-Tailed Distributions.....	22
7.19	EXCEL CONFIDENCE.T Function.....	23
7.20	Beware the Definition of a in Inverse Probability Functions	24
7.21	Interval Estimate for <i>Any</i> Normally Distributed Statistic.....	25
7.22	Population Proportion Based on Sample Proportion	26
7.23	Conditional Formatting.....	28
7.24	Confidence Limits for Population Variance and Population Standard Deviation Based on Sample Variability.....	29
8	Hypothesis Testing	1
8.1	Introduction.....	1
8.2	Are the Variances of these Two Samples the Same?	3

8.3	Levels of Statistical Significance and Type I Error: Rejecting the Null Hypothesis When it is Actually 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.....	14
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 Mean.....	20
8.14	Computing a t-test for Equality of Means without Raw Data	21
8.15	The T.TEST Function	21
9	Analyzing Relationships Among Variables	1
9.1	Introduction to Analyzing Relations	1
9.2	Cross-Tabulations (Contingency Tables).....	3
9.3	Filtering Data for Temporary Views	5
9.4	Charts for Contingency Tables.....	6
9.5	Scatterplots and the Intuitive Grasp of Relationships.....	8
9.6	Pearson Product-Moment Correlation Coefficient, r	9
9.7	Computing the Correlation Coefficient Using EXCEL.....	10
9.8	Testing the Significance of the Correlation Coefficient	12
9.9	Coefficient of Determination, r^2	13
9.10	Linear Regression in EXCEL.....	14
9.11	ANOVA with Linear Regression.....	17
9.12	Predicted Values in Linear Regression & Confidence Limits	19
10	Analyzing Frequency Data	1
10.1	Computing Expected Frequencies for Statistical Distributions.....	1
10.2	The Chi-Square Goodness-of-Fit Test	3
10.3	The EXCEL =CHISQ.TEST Function	4
10.4	Two-Way Tests of Independence Using Chi-Square.....	5
[Chapter headings for future expansion of text are omitted here]		
17	Concluding Remarks.....	1
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”² 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 *Aha!* 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.

² < <http://www.youtube.com/watch?v=e2PyeXRwhCE> >

Remember: reading about statistics must surely be the most passionately interesting and absorbing activity in your life right now ☺ 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 1st 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).

Ety~~m~~ologies

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.



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.³ 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.

Mich

M. E. Kabay
Northfield, Vermont

April 2019

NORWICH EMAIL for Norwich students, staff and faculty: <mkabay@norwich.edu>

GMAIL for everyone else: <mekabay@gmail.com>

³ Based on *Monty Python’s Holy Grail*, Scene 8. French Guard: “I fart in your general direction.”
<<http://arago4.tnw.utwente.nl/stonedead/movies/holy-grail/main.html>>

1 Introduction

1.1 About Applied Statistics

Our world and our lives are filled with counting and measuring. Applied statistics help us decide how best to count and measure, how to represent numerical information – especially information about groups – effectively and without misleading ourselves and others, and how to explore our ideas about relationships among factors that affect quantitative data. Statistics allow us to summarize large data sets in ways that let us make sense of what would otherwise be a morass of detail. They let us test our ideas to see if possible explanations are rooted in reality or whether they are impression created by the winds of chance. Statistics is often described as the art of *decision under uncertainty*.

Here are some of the many ways that statistics play a role in our studies, our work, and our daily lives.

- Business
 - A manager maintaining inventory needs to know how many products are being sold at which time of year so she can place orders before she runs out of materials.
 - What's the range of estimated sales next year that has a 95% chance of being correct?
 - What's the reliability or predictability of sales for our 14 sales professionals? Are they similar or are there differences in reliability that go beyond just random fluctuations?
 - Supervisors have to monitor the quality of their production lines and service levels to spot problem areas, inefficient processes, and people who need to improve their knowledge, skills, and performance.
 - Sales managers need to know which customers
 - Buy the most?
 - And which complain the most?
 - Are increasing their purchase levels the fastest?
 - Which salesperson has *reduced* productivity the most in the last quarter?
 - Is the reduction in sales by this salesperson due to chance alone or could there be something else going on?
 - Marketing managers ask about the millions of dollars spent on several advertising campaigns; are any of them obviously better than the others?
 - Obviously worse?
 - Are the differences just temporary fluctuations or are they something we should take seriously?
 - A health-club manager asks if a particular member's absence is just part of her normal routine or whether they should phone her to see if she would like to suspend her membership payments temporarily as part of the customer service efforts?
- Finance
 - Which of several brokerages has a reliable record of higher-than-average return on investment?
 - Is the share price for this firm rising predictably enough for a day-trader to invest in it?
 - What has our return on investment been for these two brands of computer equipment over
 - the last four years?
 - Should we pay the new premium being proposed by our insurance company for liability insurance?

- What is the premium that our actuaries are calculating for a fire-insurance policy on this type of building?
- Should we invest in bonds or in stock this year? What are the average rates of return? How predictable are these numbers?
- Which country has the greatest chance of maximizing our profit on investment over the next decade? Which industry? Which company?
- Management
 - One of our division managers claims that his below-average profit figures are just the luck of the draw; how often would he get such a big drop in profit by chance alone if we look at our historical records?
 - Another manager claims that her above-average productivity is associated with the weekly walkabout that is part of her management-by-walking-around philosophy; can we test that idea to see if there is any rational basis for believing her analysis?
 - The Chief Financial Officer is getting suspicious of these transactions in our audit trail; can we check to see if the digits conform to random expectation for their frequencies or whether they are fabricated and have non-standard frequencies?
 - The Human Resources Department wants to create a questionnaire that will help them determine where to put their emphasis and resources in internal training next year. How should we design the questionnaire to ensure the most neutral and least ambiguous formulation of the questions? How can we tell if people are answering seriously or if they are just filling in answers at random?
- Information technology
 - Which of these motherboards has the lowest rate of failure according to industry studies?
 - How long should we expect these LCD monitors to last before failure?
 - How is our disk free space doing? When do we expect to have to buy new disk drives? With what degree of confidence are you predicting these dates?
 - Which department is increasing their disk-space usage unusually fast? Is that part of the normal variations or is something new going on?
 - Which of our department has been increasing their use of printer paper the most over the last quarter? Is their rate of increase just part of the normal growth rate for the whole organization or is there anything unusual that we should investigate?
 - We have been testing three different antispam products over the last six months; is any one of them obviously better than the others?
 - Which of our programming teams has the lowest rate of coding errors per thousand lines of production code? Is the difference between their error rate and those of the other team significant enough to warrant a lecture from them, or is it just one of those random things?
 - Which programmer's code has caused the highest number of helpdesk calls in the past year? Is that bad luck or bad programming?
 - Can you identify the callers responsible for 80% of our helpdesk calls? Is there anything in common among them that would help us identify areas for improvement through increased training?
 - This student's term paper doesn't seem to have been written by him based on his previously submitted incoherent and ; can we compare the frequencies of specific words in his previous work to the frequencies in this submission to see if an investigation for plagiarism is warranted?

- Can we distinguish between the characteristics of normal software and those of malware based on what they do, how often, and how fast?
- Engineering & Production Management
 - Are these products in compliance with the regulatory requirements for our industry? How can we distinguish systemic problems from chance variations?
 - Which of these four suppliers' materials have the least variability in their products?
 - Which of these five design choices is associated with the lowest failure rate for the bridges we are building?
 - How do the following 12 options in chemical composition of a bonding material rate in terms of tensile strength? Are any of them outstandingly good or outstandingly bad?
 - Our chemical production lines are showing different results in the closeness of the products to the stated concentrations on our labels; are any of these results indicating a significant problem or are they all just random fluctuations of no significance?
 - We need to rate the rate of wear on our products according to five different factors; are any of these factors affecting each other or are they all working independently?
 - We want to create a scoring system that will give us a clear indication of whether this product is going to be acceptable to regulatory authorities or not. How do we create such a metric that will be trustworthy?
- Life in general
 - This politician claims that our taxes have been rising over the last couple of years; is that true?
 - Is using *this* brand of tooth paste associated with decreased chances of getting tooth decay compared with *that* brand?
 - Is it true that focusing one's mind on an imaginary blue sphere for 10 minutes a day is associated with raising one's IQ to 180 within a couple of weeks?
 - Does wearing a large rubber band around one's abdomen really seem to be associated with weight loss of 30 to 80 pounds in a month?
 - Is taking vitamins A, B, C, D, E, and K every day associated with reductions in the risk of blindness, cancer, cataracts, joint degeneration, liver disease, and laziness?
 - Does listening to two different sounds at the same time really relate to changes in brainwaves that make us feel calmer and be smarter?
 - These people argue that illegal immigrants are associated with lower-than-average crime rates – is that a real phenomenon or just the luck of the draw?

1.2 Computations in Applied Statistics

The first question before turning to the specific software package is why students should be using computer programs for their first course in statistics instead of working through the problems using equations and calculators. From my perspective as someone who actually did that in the 1960s, we might as well ask why the students shouldn't be moving little stones around (calculi) to do their calculations and then chiseling their results on stone tablets. I find no pedagogic benefit whatsoever in forcing students to waste their time in manual calculation when there is a wealth of excellent statistical packages available and in use in the real world.

I emphasize to students in my initial lecture that I am much less interested in any one statistical package than I am in the skills I want them to develop in *how to learn to use any statistical package*. Students should learn to use help facilities and documentation included with their program, online tutorials and discussion groups about the software, and trial-and-error experimentation as they master *any* new program.

1.3 Why EXCEL as Opposed to Other Statistical Packages?

- Availability: the program is almost universally available in universities and business; in addition, MS-Office software is available at deep discounts for students to install on their own computers.
- Simplicity and familiarity: many (not all) students already have at least a passing knowledge of EXCEL and at worst, are familiar with the style of its user interface. In contrast, many statistical packages have user interfaces that are radically different in conception and detail from the office software with which students are already familiar.
- It's easier to learn a new statistical package once one becomes familiar with a simpler one.

1.4 Learning EXCEL

Throughout this text, EXCEL functions are used and sometimes illustrated for the statistical methods under discussion. The only way to learn how to use this tool is to *use the tool*.

Some key resources for learners:

- The widespread use of EXCEL has led to a wealth of free tutorials online.⁴
- Conrad Carlberg's *Statistical Analysis: Microsoft Excel 2010* provides a superb reference guide in paper and electronic versions for modest prices.⁵ A sample chapter is available online.⁶
- Students have access to their own computers and inexpensive student licenses for EXCEL.⁷
- All the computer lab computers provide access to EXCEL.
- Help is available within the program by clicking on the Help  symbol or by pressing function key 1 (F1) on the Windows keyboard.
- Using the File | Help menu (Figure 1-1) brings you to a menu of documentation and online demos and tutorials that demonstrate the fundamentals of using the graphical user interface.

⁴ Type *excel tutorial* into the search field of GOOGLE for a wealth of options. To the degree possible, use the tutorials available from educational institutions (.edu) to avoid getting tangled up in commercial sites that could spam you later. Inexpensive self-contained training courses on CD-ROM for more advanced functions include the "Excel Advanced Interactive Tutorial" for PCs, which is available for download at \$14.97 from <<http://www.deluxetraining.com/EXCEL-Tutorial.html>>. [Note: the author has no relationship whatever with these providers.]

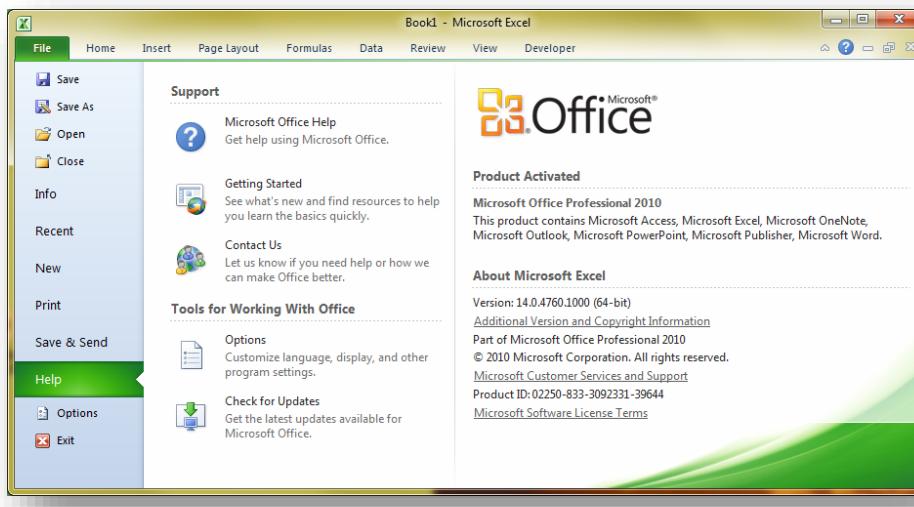
⁵ (Carlberg 2011)

⁶ See <<http://www.quepublishing.com/articles/article.aspx?p=1717265&seqNum=3>>.

⁷ Norwich University has registered with journeyEd.com for low student prices: go to <<http://www.journeyed.com/students>>.

- The blue Help symbol brings up a complete reference manual with search capability. Typing “get started with excel points” links to several useful tutorials.

Figure 1-1. File | Help menu in Excel 2010.



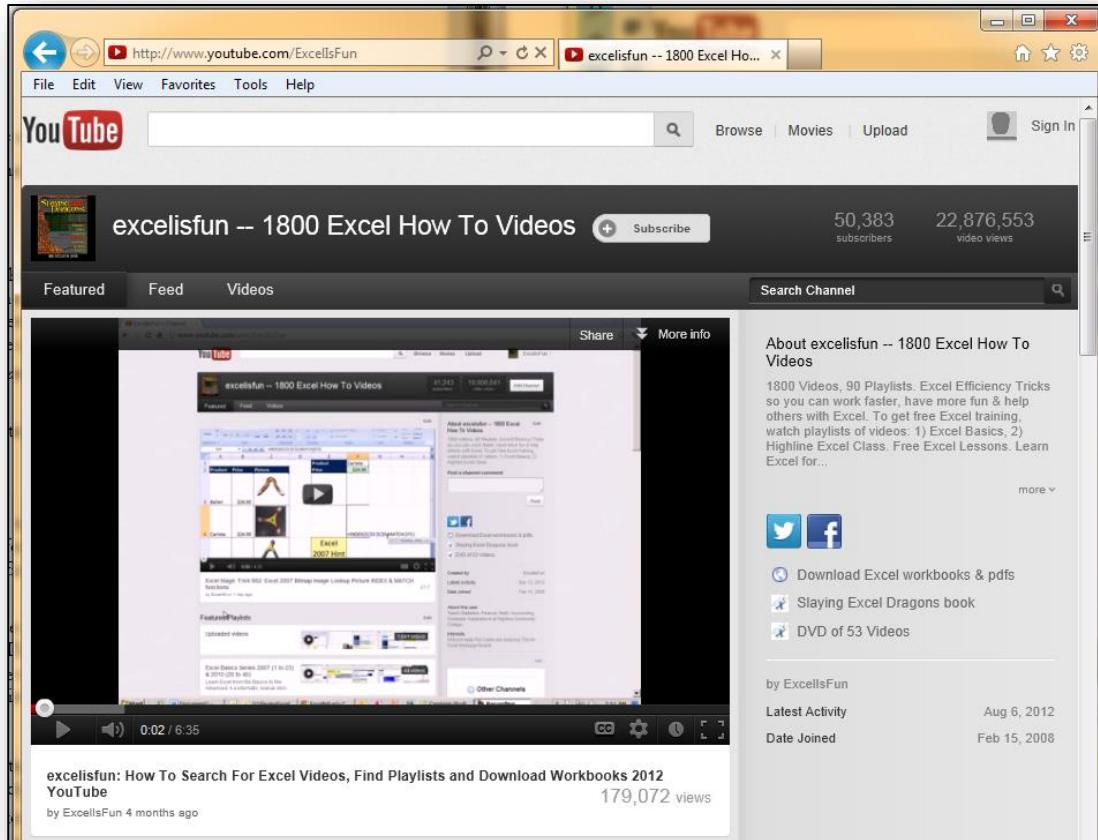
- In the QM213 course, EXCEL is integrated into all the homework assignments:
- All the homework assignments are provided in EXCEL workbooks corresponding to the chapters.
- Each EXCEL workbook has several worksheets with different problems or types of problems.
- Instructions for using EXCEL in solving statistical problems are included in this textbook; some are presented in class as demonstrations or, once students gain more experience, are available on demand from the instructor.

In the QM213 course, EXCEL is integrated into all the homework assignments:

- All the homework assignments are provided in EXCEL workbooks corresponding to the chapters.
- Each EXCEL workbook has several worksheets with different problems or types of problems.
- Instructions for using EXCEL in solving statistical problems are included in this textbook; some are presented in class as demonstrations or, once students gain more experience, are available on demand from the instructor.

- An extraordinary collection of instructional videos about EXCEL running under Windows is available free online at <<http://www.youtube.com/ExcelIsFun>> as shown in the screenshot in Figure 1-2. These videos are like having a personal tutor for every aspect of EXCEL under Windows.

Figure 1-2. Excel Is Fun YouTube Channel <<http://www.youtube.com/ExcelIsFun>>.



- The subseries of particular interest for students in QM213 are
 - Excel Basics Series 2007 (1 to 23) & 2010 (20-40)
 - Highline Excel Class Beg – Adv (Complete Class – 59 videos)
 - Excel 2010 Statistics Formulas Functions Charts PivotTables (91 videos).

Many associated spreadsheets for these lectures are available for free download
<<https://people.highline.edu/mgirvin/ExcelIsFun.htm>>.

For students who use Macintosh computers, Microsoft.com has several hundred instructional materials online; use *Excel for Mac* as the search string or go directly to <<http://tinyurl.com/9dvy2ur>>.⁸

- Students are encouraged to contribute additional or specific EXCEL training resources (articles, books, video) using the NUoodle classroom discussions.

⁸ TinyURL.com converted a 107-character URL into a link with 7 characters after the standard <http://tinyurl.com/> strong.

1.5 Long-Term Goals: A Personal Perspective

I despise rote memory work. I tell students not to memorize; it doesn't do them any good in the long run. Cramming meaningless detail into one's brain simply causes most of the material to ooze out the cracks (what a distasteful image). Much more important than merely passing a quiz this week and forgetting its contents next week are the following fundamental abilities, which can last a lifetime:

- *Long-term learning:* Integrating new knowledge into a solid network of associations with practical examples, easily-accessed principles for application, and experience with many exercises that convert new information into permanent knowledge.
- *Critical thinking:* Students should be able to look at poorly conceived statistical reports with a jaundiced eye; they should be able to ask serious questions about the methodology used in gathering data and the methods of analyzing and representing them.
- *Ability to learn at will:* This introduction to statistical thinking should help students learn any statistical methods they need in the future.
- *Applications to current and future studies:* Students should be well prepared for more advanced courses in applied statistics and for courses using statistics in their areas of study. I've already received many reports from students and from my fellow statistics professors that students who do well in QM213 are doing well in the more advanced statistics course and in finance and marketing courses as well.
- *Integration into core thinking in life and work:* The rational approach to thinking about quantitative data should help students resist propaganda and flimflam.

Finally, I want students who invest the time and effort to master the skills presented in this text to *enjoy* their classes in applied statistics. Seeing students smiling during and after a statistics lecture is a wonderful experience for a teacher, considering the scowls usually associated with any statistics course.

1.6 Using NUoodle

QM213 uses the electronic teaching platform NUoodle, the Norwich University implementation of Noodle. Tutorials on using NUoodle are available online once you have logged on to the system.

The NUoodle classroom for QM213 includes

- Links to resources such as course description, syllabus, and this textbook
- Sections corresponding to the weekly assignments
- All homework assignments as XLSX files
- Discussion groups
- All quizzes and examinations.

INSTANT TEST Page 1-8

- (1) Explore several of the Excel resources listed above and report on what you find by posting comments about which ones you liked (and why) and which ones you disliked (and why).
- (2) Explain to yourself or to a buddy why you should integrate the subject matter of your course instead of simply focusing on maximizing your grade (even if you forget everything soon after the course ends).
- (3) Talk to students who have completed QM213 and ask them about their impressions of the course.

1.7 SQ3R

I recommend to students that they use the SQ3R methodology throughout their work to increase the effectiveness of their studies. The acronym stands for “Survey, Question, Read, Recite, Review” and describes a systematic approach to making full use of their time when reading any technical material. There are many descriptions of this method on the Web; they can find hundreds by entering “SQ3R” into any search engine; most university learning centers will also be able to provide resources to learn about and apply SQ3R. A brief summary of the method follows and is written as instructions to the student:

1. Survey

Survey the document: scan the contents, introduction, chapter introductions and chapter summaries to pick up a general impression of the text. Leaf through the entire text quickly (this may take you an hour) and let your eyes glance over each page. Allow yourself to look at pictures or diagrams and to read any figure captions or highlighted words. Let yourself become curious. Make notes on any questions that occur to you (see below in the Question section). This is not reading in any detail: you are just starting to build a framework of associations that will help you assimilate the information you will study.

2. Question

Throughout your survey and as you read the material, write down questions that come to mind. Not only are you maintaining an active stance in your studies as you do this, but you are building a stockpile of test questions that will help you check your own comprehension later.

3. Read

The trick here is to read one paragraph at a time. The paragraph is the logical unit of thought (if the writer has any sense); you can best master the information by reading one paragraph and then immediately moving to the Recite phase.⁹

4. Recite

No, this is not memorization. You look up from the paragraph you have just read and ask yourself, “What have I read?” Just summarize the main point(s) in your own words. The Read-Recite cycle turns passive recognition of knowledge (“Oh yeah, I read that”) into active knowledge (“This is what it means”). You need active knowledge to be successful in any course.

5. Review

Consolidate your new knowledge. Review the main points you’ve learned after every study session. Check your own or the teacher’s review questions to test and solidify your knowledge. Go back to your study materials and fill in the holes if necessary. Then later, at the end of the day and the end of the week, repeat your review to help solidify your memory and to continue the process of making the knowledge your own.¹⁰

INSTANT TEST P 1-9

Close this page and explain to yourself or a buddy every step of the SQ3R method and why it is helpful.

⁹ (Adler and Van Doren 1972)

¹⁰ For additional guides to techniques that may help students, visit <<http://www.mekabay.com/methodology/index.htm>> and download “Computer-Aided Consensus™”, “Frequently Corrected Errors”, “On Writing”, “Organizing and Safeguarding information on Disk”, “SQ3R” (the source for the material above), “Tips for Using MS-WORD 2007”, “Tracking Changes in MS-Word 2003”, and “Understanding Computer Crime Studies and Statistics.”

1.8 Counting and Measuring

Evidence from biological research suggests that bees, primates, and dogs can count; evidence from anthropology and archeology suggests that human beings have been counting for at least hundreds of thousands of years. Early counting probably involved using markers such as stones (the Latin for little stone is *calculus*, hence our verb *calculate*) to keep track of how many objects or groups of objects had been noted. Children in numerate cultures learn to count properly between the ages of three and five years. Although some cultures reportedly limit the complexity of their enumeration to roughly “one, two, many,” modern society has extended its range of counting into unimaginable reaches such as the googol (10^{100}) and the googolplex (10^{googol})¹¹ (but not the Googleplex¹², which is something else altogether). To put these numbers in perspective, there are roughly 10^{80} subatomic particles in the entire known space-time continuum (excluding any parallel universes or the amount of money being earned by Google).

Measurement consists of *counting* the number of *units* or *parts of units* displayed by objects and phenomena. The development of science has been reflected to a large extent in the availability of an ever-widening array of units of measurement; for example, we can measure the usual attributes such as size, mass, and force but also hidden aspects of reality such as up/down, charm/strange and top/bottom aspects of quarks. Sometimes we invent measures that turn out to be dangerous: the use of *derivatives* in the real-estate stock market is usually thought to have been at the root of the financial collapse of the late 2000s.¹³

Some scales are attempts to quantify reality by assigning numerical values to subtle aspects of the world; the classic example of quantifying the hard to quantify is the Likert Scale, invented by Rensis Likert (1903-1981), a psychologist who assigned numerical values to assertions about human attitudes. All of us are now familiar with questions that have answers in the form

- 1 – strongly agree;
- 2 – agree,
- 3 – neutral;
- 4 – disagree;
- 5 – strongly disagree.

The statistical methods used to obtain, summarize and analyze such data are widely used.¹⁴ However, simply adding up scores in the expectation that the totals will be meaningful is an error that has demolished many student (and even professional) research projects. You can add some data but you can't add others, and Likert scale results *cannot usefully be added up and averaged*.

Categorizing (classifying) aspects of our world consists of labeling what we encounter with descriptions that define classes; for example, we can classify students by

- Their expected year of graduation (class of 2028, class of 2029...) or by
- The majors in which they are enrolled (computer security & information assurance, construction management, business management, ...), or by
- The level of their degree (undergraduate, graduate) or by
- Their enrolment status (Corps of Cadets, civilians).

The details of how we think about and use information are fundamentally important in statistics. The concepts discussed in the next sections affect how we choose to record, represent and analyze the results of our investigations.

¹¹ (Wolfram Mathworld 2012)

¹² Google's corporate HQ at 600 Amphitheatre Parkway. Mountain View, CA 94043, USA

¹³ (Washington's Blog 2008)

¹⁴ (Trochim 2006)

1.9 Variables

The results of counting, measuring and classifying are data. *Data*, the plural of the Latin word *datum*, meaning an item of information, are the raw material of statistical analysis.¹⁵ Throughout this course, you will have to recognize and distinguish between two types of data, corresponding to what you count and what you measure: *categorical* or *qualitative* data versus *quantitative* data.

- Examples of qualitative data
 - Color (e.g., red, orange, yellow, etc.)
 - Tone (e.g., middle-C, b-flat below middle-C)
 - Timbre (e.g., oboe-sound, flute-sound, drum-sound)
 - Shape (e.g., round, square, tetrahedral, etc.)
 - Preference for a particular movie (e.g., like, neutral, dislike, etc.)
 - Type (e.g., wool, cotton, plastic; or complaint vs praise, addiction vs habit; etc.)
 - Origin (e.g., endogenous vs exogenous; local vs foreign)
- Examples of quantitative data
 - Primary wavelength of light reflected by an object (e.g., 650 nm for a “red” object) Primary frequency of light reflected by an object (e.g., 640 THz for a “blue” object)
 - Waveform of a trumpet sound expressed as Fourier transforms of a sonogram
 - Length of sides of a triangle
 - Number of people expressing a particular preference for a movie
 - Numerical representation (e.g., a Likert scale) of a feeling (e.g., 0 = no pain, 1 = barely noticeable pain, 2 = slightly annoying pain, ... 10 = “Angel of Death please take me now.”)

INSTANT TEST Page 1-11

Which of the following is a qualitative variable and which is a quantitative variable?

- (1) Saying that a chocolate bar “tastes good” or “tastes bad.”
- (2) Weighing a chocolate bar in grams.
- (3) Evaluating the grams of fat per gram of chocolate bar.
- (4) Counting the number of students who say that a chocolate bar tastes good and the number of students who say that a chocolate bar tastes bad.
- (5) Describing a fellow student as “attractive” and another one as “not attractive.”
- (6) Classifying a professor as “so boring I have to text my friends 38 times per period to stay awake” or as “interesting enough to stop me from texting in class.”

¹⁵ Although technically, one should write, “Data are....”, almost everyone today writes, “Data is....” Students may use either form without fear. Don’t be surprised to hear the (old) professor say “....data are....”

1.10 Tables

Figure 1-3 is a simple table (a matrix with labels for the *rows* down the side and those for the *columns* across the top) showing some (invented) qualitative and quantitative data. Such a collection of information is often the result of a *study* (sometimes called a *case study* or, if a change was imposed on the situation, a *trial*).

In Figure 1-3 the focus of the study is on *different characteristics* of the five divisions in the company. Division can be called the *identifier* and the information about the different divisions can be called *observations*.

Figure 1-3. Observations from research study on Urgonian Corporation divisions.

Division	MBWA Used? (Y/N)	Total Employees in Plant	Team Rank in Company Soccer League	Average Monthly Profit (U\$) per Employee	% HelpDesk Calls attributed to Lack of Training
Akron, OH	N	4,392	5	\$590	13.2%
Bayreuth, Germany	Y	3,054	2	\$436	4.1%
Canberra, Australia	Y	2,855	3	\$631	3.2%
Kuala Lumpur, Malaysia	N	9,218	1	\$755	16.9%
Oakland, CA	Y	1,627	4	\$821	3.7%

The different observations about each of the specific cases of the identifier are called *variables*; in this case, we have a total of six variables because the identifier is itself a variable. All the variables are listed in columns in this table. Two of the variables are *qualitative* (sometimes described as *categorical* because values are categories) and four of the variables are *quantitative*. Quantitative variables may be measured, counted, or ranked.

Qualitative variables in our study:

1. The use of MBWA (management by walking around) – qualitative
2. The name or location of the division – qualitative.

Quantitative variables:

3. Total employees in plant – quantitative (counted)
4. Team rank in the company soccer competitions – quantitative (rank order)
5. Average monthly profit per employee in US dollars – quantitative (measured)
6. The percentage of HelpDesk calls traced to a lack of training – quantitative (measured)

INSTANT TEST Page 1-12

Suppose you were looking into the use of alcohol by students in different residence halls of your university. If you recorded "YES" and "NO" as the only data for alcohol use, would alcohol use be a quantitative variable or a qualitative variable? What if you recorded the number of ounces of alcohol consumed per week? What would that be?

1.11 Choosing a Table Layout

Is there any significance to whether the identifier is placed down the left-hand column or across the top row?

Nothing stops us from representing the same data we've been discussing in a table where the divisions are arranged as the heads of columns instead of the labels for rows. In Figure 1-4 each observation corresponds to one column instead of one row – the specific values of the five variables for each division are written out across the top row as the heads of the columns.

Figure 1-4. Divisions across the top header.

Division	Akron, OH	Bayreuth, Germany	Canberra, Australia	Kuala Lumpur, Malaysia	Oakland, CA
MBWA Used? (Y/N)	N	Y	Y	N	Y
Total Employees in Plant	4,392	3,054	2,855	9,218	1,627
Team Rank in Company Soccer League	5	2	3	1	4
Average Monthly Profit (U\$)	\$590	\$436	\$631	\$755	\$821
% HelpDesk Calls attributed to Lack of Training	13.20%	4.10%	3.20%	16.90%	3.70%

Orientation of the table is a subtle issue; there is *nothing wrong with either format* – they are both acceptable in practice. However, the decision about which way to arrange data may depend on our needs. Sometimes if the names of many variables are longer than the names of the elements, we may want to give them more room by putting them in a *single* wider column (that is, listing them vertically down the left side of the table) instead of using up valuable space on a page with many wide columns.

However, the most important issue in practice is clarity: in general readers tend to look across each row from left to right and then go on to the next row. Thus some readers seeing Figure 1-3 may consciously or unconsciously expect the focus of the discussion to be on the each of the geographic locations (*Akron OH* first, then *Bayreuth* and so on) with details (MBWA Used, Total Employees...) discussed for each plant.

In contrast, the first impression created by Figure 1-4 may be that the discussion is going to be about the pattern of observations of the different variables for each location in turn: *the state of MBWA Used for Akron*, then in *Bayreuth*, and so on; that discussion would then be followed by the examination of the *Total Employees in Plant* for *Akron*, for *Bayreuth* and so on.

Nonetheless, either form is acceptable in practice. These issues are relatively subtle aspects of professional presentation and even experienced writers and presenters may differ in their opinions on the subject.

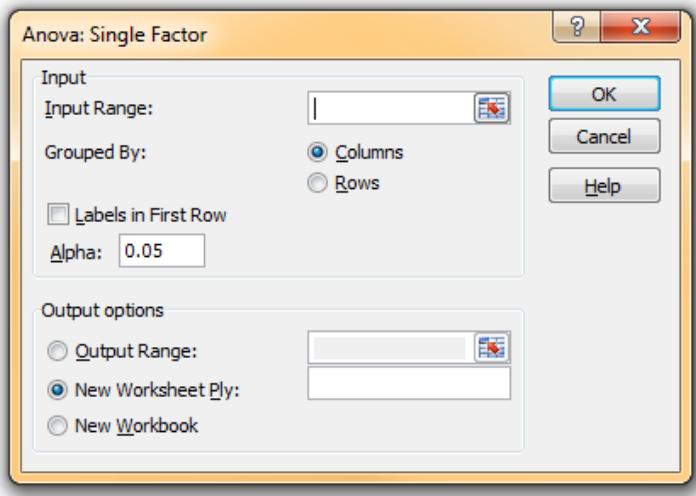
INSTANT TEST P 1-13

Show several fellow students who have not taken statistics courses how they look at Figure 1-3 and Figure 1-4. Which way do they read the data (ask them to talk out loud as they look at the tables). Report on your findings in the QM213 discussion group for this week.

1.12 Transposing Rows and Columns

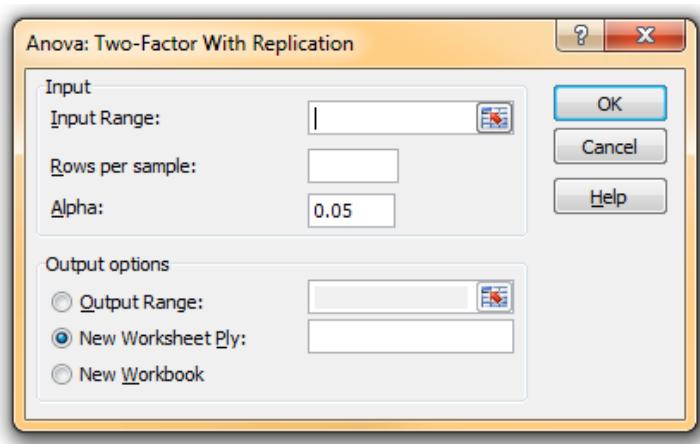
In EXCEL and other statistical packages, an additional consideration beyond simply optimizing appearance is whether a particular statistical function or routine requires a particular arrangement of data to function properly. Such constraints are always defined in the accompanying Help texts for the function. For example, Figure 1-5 shows a statistical routine called **Anova** (usually printed as *ANOVA* in most discussions¹⁶) – it stands for *ANalysis Of VAriance*) that is widely used to find out if groups of observations have similar characteristics).

Figure 1-5. ANOVA function that accepts data in rows or in columns.



In contrast, Figure 1-6 is a kind of ANOVA function in EXCEL that absolutely requires data to be arranged with the names (identifiers) of the groups in question across the top of the table and the individual data arranged vertically in columns with a specific number of rows for each group.

Figure 1-6. ANOVA that requires data in columns.



¹⁶ Nothing to do with *new*; *nova* is a Latin form of its word for *new* (e.g., *nova* for a new star, *stella nova*) but *ANOVA* is just an acronym.

If you decide to switch the orientation of a table you can copy it and then paste it into a new position using the CTL-ALT-V key combination or the Paste Special function button and checking the Transpose option as shown in Figure 1-7.

To transpose a table, highlight the whole table and copy it (CTL-C or Copy) as shown in Figure 1-9. Highlighted table for pasting (note the highlighting and the twinkling outline on your own screen if you try this yourself). Then move the cursor to the appropriate new position for the transposed table and press CTL-ALT-V. The results are shown in Figure 1-9.

Figure 1-9. Highlighted table for pasting.

	A	B	C	D	E
1	Table 1	Column B	Column C	Column D	Column E
2	Row 2	b2	c2	d2	e2
3	Row 3	b3	c3	d3	e3
4	Row 4	b4	c4	d4	e4
5					

Figure 1-7. CTL-ALT-V / Paste Special dialog.

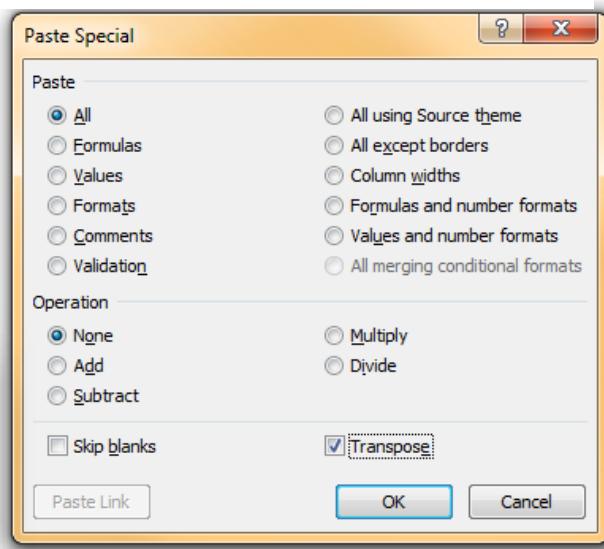


Figure 1-8. Transposed table.

	A	B	C	D
1	Table 1	Row 2	Row 3	Row 4
2	Column B	b2	b3	b4
3	Column C	c2	c3	c4
4	Column D	d2	d3	d4
5	Column E	e2	e3	e4
6				

If your table has borders, transposing the cells will likely result in a mess, as shown in Figure 1-10. You can just erase the borders if that happens by mistake and then put in new ones.

Figure 1-10. Transposed table with bad borders.

	A	B	C	D
1	Table 1	Row 2	Row 3	Row 4
2	Column B	b2	b3	b4
3	Column C	c2	c3	c4
4	Column D	d2	d3	d4
5	Column E	e2	e3	e4
6				

To erase borders, highlight the table and click on the pull-down menu to select No Border: Then apply appropriate borders as you see fit.



1.13 Types of Variables

Earlier (§1.9) you read about types of variables. It's important to think about the type of variable we are working with because different statistical methods apply to different kinds of information. For example, we can easily think about the average height of our fellow students, but it's inappropriate to try to define an "average" that corresponds to their preference in reading materials. We use other methods such as frequency distributions for such data.

A fundamental distinction is between *quantitative* and *qualitative* variables, as first discussed in §1.9.

1.14 Qualitative / Categorical Data and Nominal Scales

Figure 1-3 and Figure 1-4 show variables that don't have measurements at all: they are *labels* or *categories*. The use of MBWA (YES / NO), for example, is called a *nominal* or *categorical* variable and it uses a *nominal scale* of measurement.¹⁷ These variables don't have numerical values assigned to them; there is no way of counting them or measuring them.

Examples of qualitative variables come up constantly in our experience. For example, you may be studying

- The presence or absence of steel reinforcement bars in concrete,
- The type of wing in aircraft (fixed, rotatory),
- The country of origin of immigrants (Afghanistan, Albania, Algeria, Angola ... Venezuela, Vietnam, Yemen, Zambia, Zimbabwe),
- The manufacturer of cars (Acura, Alfa Romeo, American Motors, Aston Martin ... Toyota, Triumph, Vauxhall, Volkswagen, Volvo), and
- The flavor of *Bertie Bott's Every Flavour [candy] Beans* (including banana, blueberry, booger, cherry, earwax, jelly slugs, licorice, liver, rotten egg, sausage, soap, and vomit).¹⁸

Qualitative variables do not necessarily have meaningful relationships among specific values; thus there is no limitation on which order one could list the divisions in our tables. Figure 1-3 and Figure 1-4 happen to list the divisions in alphabetical order by city, but no one should assume that the order implies anything else about the elements. One could equally well have listed the divisions according to some other arbitrary scheme such as alphabetical order by country or simply personal preference.

1.15 Quantitative Data

Most of the statistical work we will be doing concerns counting and measuring. These *quantitative* variables are familiar to all of us. What you count or measure are the variables; the results of your counting and measuring are the data. Examples include number of offspring, income, expense, mass, weight, height, tensile strength, concentration of chemical constituents, rank in a contest, and so on.

INSTANT TEST P 1-16

Look up a statistical report on something you find interesting (science, sports, games, weather, politics, networks, security, engineering, construction management...) and analyze what kinds of variables you find in them. Report on your findings in the NUoodle discussion group for this week.

¹⁷ From *nomen*, the Latin word for name

¹⁸ At the time of writing, these Hogwarts confections are available for purchase and, er, consumption from the Jelly Belly Shop. <<http://www.jellybelly.com/Shop/ProductDetail.aspx?ProductID=98101>>

1.16 Discontinuous / Discrete Variables

Another distinction involves the measurement scale for quantitative data. Consider the *number* of employees per plant: clearly the only values acceptable are whole numbers (*integers*). We call these data *discrete*¹⁹ or *enumerated*. Many variables are described using discrete data; any time the underlying phenomenon involves distinct units or steps, the data will be discrete. Thus the *number of cylinders in a motor*, the *number of members on a sports team*, the *number of clients supported by a broker*, the *number of lawsuits* launched by the broker's clients after they find out what the broker did to them – all of these are discrete variables.

Certain kinds of statistical methods work effectively with discrete variables; for example, if customers buy different kinds of product in different demographic groups or geographic areas, statisticians are likely to compute frequency distributions of the number of purchases of each type of product in each subgroup. They can then use a *test of independence* of the frequency data to see if the differences in the *samples* reflect differences in the *populations*.²⁰

1.17 Continuous Data

If something can *in principle* assume any possible value between two limits, the variable is described as continuous. Thus the *average monthly profit* and the *percentage of HelpDesk calls attributed to lack of training* are both considered continuous variables when there are many observations.

Students sometimes express surprise that summary statistics such as averages (you know about those – more details later) for discrete data often have fractional values; however, there would be nothing unusual about defining the *average number of employees per plant* in our tables as 4,229.2 even though there is no such thing as 0.2 of an employee. Such values are simply the result of the definitions and computations which generate continuous variables even though the *raw* data are discrete.

Some statistical methods apply only to continuous data; for example, the *analysis of variance (ANOVA)* assumes a continuous measurement scale. If we make the mistake of applying ANOVA to frequencies, the arithmetic may work out, but the results will not mean very much because the assumptions of ANOVA will not have been met. In such cases, we may be misled into erroneous conclusions about the underlying phenomena we are exploring.

INSTANT TEST P 1-17

If you count the number of fingers and toes you have, why is that variable discontinuous when you measure your weight by counting the number of pounds you have amassed in a lifetime of self-discipline and continuous self-monitoring? Why is weight viewed as a continuous variable? Explain this contradiction as if speaking to an intelligent 10-year old.

¹⁹ Don't use the spelling *discreet* for this concept – *discret* means *tactful, good at keeping secrets, unobtrusive, or modest*.

²⁰ We discuss samples and populations in great detail later in the text. A population is all possible members of a group. Thus if we are studying characteristics of current full-time students at Norwich University, then the population is all of those students. A sample is part of a population or a specimen for analysis. Following the idea of defining all the current full-time students at Norwich University as the population of interest, a group of students in the QM213 course might be considered a sample. A random sample is a sample in which every member of the population has an equal chance of being selected for the sample – thus the QM213 students cannot be a random sample of all the students currently enrolled in the University. The notion of a random sample is fundamentally important in statistics.

1.18 Interval Scales

If the measurement scale defines a constant quantitative difference between values that appear to be sequential, it's called an interval scale. For example, the Celsius, Fahrenheit, and Kelvin temperature scales all define fixed differences in temperature between values. Thus the Celsius scale defines 100 units (degrees) between the temperature of pure water at the freezing point at one atmosphere of pressure and its temperature at the boiling point. A 20 degree difference in temperature (e.g., 40C vs 60C) is defined as equivalent to any other 20 degree difference (e.g., 400C vs 420C). The same rules apply to the Fahrenheit scale (with the freezing point of water defined as 32F and its boiling point as 212F. Again, a 20 degree difference is defined as equivalent to any other 20 degree difference (e.g., 400F vs 420F). The same principle applies to the Kelvin scale.

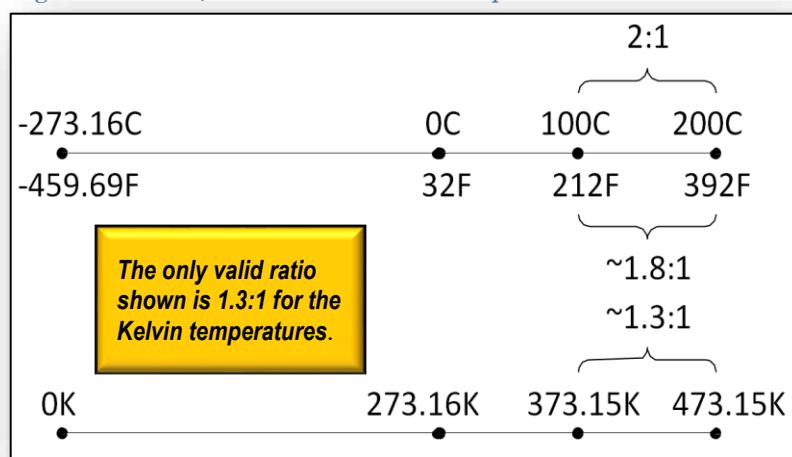
Interval scales, because of their equal divisions, *can* be subjected to addition and subtraction. For example, if one measures the temperature of two objects as 15F and 20F, it is legitimate to calculate that their *average* temperature was $(15+20)/2 = 17.5\text{F}$.

1.19 Ratio Scales

Another characteristic of some interval scales is that they have a *meaningful zero*. A zero value on this scale implies absence of whatever is being measured or counted. For example, if we count When we have 0 donuts, we have no donuts, but in contrast, when we have a Scholastic Aptitude Test (SAT) score of 200, it's the lowest adjusted score, but that does not imply that a student got zero correct answers.²¹

Consider the three temperature scales introduced above. Only the Kelvin scale sets its zero value at the theoretically complete absence of what it measures. The Celsius scale of temperature defines the temperature of freezing pure water at atmosphere pressure as 0C, but the complete absence of molecular motion theoretically occurs at absolute zero, which is -273.16C or -459.6F. Thus the Celsius scale is *not a ratio scale* because we can't compute the ratio of molecular motions from different temperatures. A temperature of 30C is *not* twice as hot as a temperature of 15C. The same observation applies to the Fahrenheit scale, since 60F is *not* twice as warm as 30F in any sense. In contrast, the Kelvin scale, which defines its zero at a value (absolute zero) that corresponds to an underlying aspect of reality – molecular motion – really is a ratio scale: there is genuine meaning to the assertion that a temperature of 120K is *twice as warm* as one of 60K. Figure 1-11 below shows how meaningless the ratios are in interval scales that are not ratio scales.

Figure 1-11. Kelvin, Celsius and Fahrenheit temperature scales.



²¹ (CollegeBoard 2012)

A letter to the editor that dealt with this issue of paying attention to whether a measurement uses a ratio scale or merely an interval scale was published in 2010 in a well-known British newsmagazine:

In her letter from Madagascar (2 July) Georgina Kenyon reports: "According to the World Bank, in the past 50 years the country has seen a 10% increase in temperature and a 10% decrease in rainfall." The latter is numerically possible but the former is numerically meaningless since any absolute change in temperature will represent different percentages according to the temperature scale being used (as they differ with respect to their origins). Hopefully the World Bank – if correctly quoted – is more numerate in other areas. – Philip Lund, Nantwich, Cheshire, UK.²²

All *ratio scales* must include a zero value showing complete absence of whatever is being measured, not simply an arbitrarily chosen point on the scale.

1.20 Ordinal Scales: Ranks

It's important to repeat that not all quantitative data are on ratio scales. For example, *rank orders* are not even interval scales, let alone ratio scales: the rifle-shooting score of the top and second-ranking competitors are not guessable simply from their ranking, and the ratio of their scores cannot be determined simply from their ranks. Similarly, the scores offered by judges at figure-skating events don't automatically scale to comparable numerical results: someone with a 10 score is in no sense "twice as good" as a competitor with a 5 score. Indeed, the ratio of scores will differ depending on the quality of the athletes: score of 10 and 5 mean radically different degrees of difference in skill in a competition among five-year-old figure skaters than in a competition among Olympic-class skaters. Depending on the competition, the performances of the number one and number two skaters might be theoretically evaluated as 9.999 and 9.998 on a quantitative scale in a class of experts but as (say) 6 and 2 in a class of amateurs.

To reiterate, ordinal scales have no implication of *equal* distance, intensity, or any other measure among the numbers. They are not interval scales. If we define rank 1 as the top and rank 5 as the bottom, there is no implication that the teams are spaced equally from top to bottom as defined by their ranks. Thus a difference between first place and second place in a contest cannot legitimately be interpreted as somehow equivalent to a difference between second place and third place.

Furthermore, ordinal scales are not ratio scales: there is no implication of a meaningful zero in an ordinal scale. First place, second place and so on do not imply a "zeroeth rank." Because ordinal scales are not ratio scales, we can never meaningfully compute *ratios of ranks*. No one should expect that if the German team is ranked second in a contest and the Malaysian team is first, then the German team is "half as good" as the Malaysian team simply based on rank – or that any meaningful ratios could be ever be deduced from any ranks.

INSTANT TEST P 1-19

An inexperienced intern is given the following information about the total yearly sales of five salespeople in three years of sales (\$000s). Calculate the three-year averages of the total sales for each these salespeople and discuss why the rank of those averages conveys less information than the actual averages. Post your discussion in the NUoodle discussion group.

Year:	Adam	Betty	Callie	Donald	Ephraim
2071	391	292	623	387	176
2072	408	293	652	412	191
2073	441	395	760	435	250

²² (Lund 2010)

1.21 Identifying the Type of Variable Really Matters to You

These subtle categories among variables really *do* matter to anyone hoping to use statistical methods usefully because the tools for statistical analysis often depend on precisely what kind of measurement is being analyzed. For example, there are assumptions about the nature of variables underlying popular methods such as ANOVA (analysis of variance) and regression (curve fitting); if those assumptions don't apply, we may have to modify the measurement scales (applying *transforms*) or use different methods (e.g., *non-parametric tests*).

When the nature of the underlying variables is misunderstood, calculations using inappropriate methods may give the *illusion* of being meaningful, but they may mislead the user into conclusions that are *wrong*. Effective use of statistics cannot be a charade or a shadow-play in which we merely pretend to be applying reason to solving problems by going through the motions of analysis without checking our assumptions.

Real world consequences can result from using the wrong statistical method for the type of variable:

- Companies may spend money on the wrong products,
- Buildings may collapse,
- Marketing campaigns may fail,
- Government programs may spend money on the wrong problems using the wrong solutions,
- Ships may sink,
- Doctors may prescribe the wrong pharmaceutical and
- People, other living things and the entire biosphere may suffer.

Developing the habit of ensuring that one's calculations are correct before relying on them or before submitting them to others for their use can be a lifelong contribution for ensuring a successful career.²³

²³ **Personal notes on the value of knowing statistics:** As I mentioned in my biographical sketch in the Background section at the start of the text, one of the reasons I was accepted into a doctoral program in invertebrate zoology at Dartmouth College was my interest in statistics; I was able to make some constructive suggestions to my research director in my first interview with him (even though he was astonished at my chutzpah). With his support, I was granted a full waiver of fees (\$10,000 a year at that time) as a Teaching Fellow (three years) and as a Research Fellow (one year) for my entire time at Dartmouth College – and an additional stipend of \$11,000 per year to cover my living costs! In today's dollars (~5:1), knowing about applied statistics (and having the nerve to offer improvements to a distinguished scientist's statistical methods) earned me something in the order of \$400,000 of value.

Those suggestions were specifically related to the type of data being analyzed: he had been counting the number of rotifers (little water animals) classified into five categories according to their shapes. He would then apply ANOVA (analysis of variance) techniques to compare the distributions of shapes in different treatments. Unfortunately, ANOVA is predicated on the notion that we are analyzing continuous variables. It does not work on frequency data of the kind my director was using. He had actually been bewildered by why his statistical results were not indicating the kinds of obvious differences that were apparent in the data. Applying appropriate methods for ordinal variables using analysis of frequencies solved the problem: work that had incorrectly been judged non-significant because of the erroneous choice of methods was finally – correctly – evaluated as highly significant.

A few years later, I went for interviews at a well-known university and sat with every faculty member to discuss their research; in most cases, I was able to suggest methods that could help them improve their statistical analysis. After I hauled out my personal computer and did an analysis on the spot for him, one professor actually dragged me excitedly to his colleague and practically shouted, “He – he – he solved a problem in my research I’ve been struggling with for the last three years!” I got the job offer (but turned it down when they offered me barely more than a graduate-student’s stipend).

Another personal story is about the daughter one of my family friends. During her university studies, she did exceptionally well not only in her studies of social psychology but in particular in her applied statistics courses. I urged her to pursue her interest in statistics as a second string to her bow; every psychology department would be even more interested in hiring her as a professor if she could also teach the statistics courses. Sure enough, she is an award-winning professor of social psychology and has been enjoying teaching experimental design and statistical analysis since 1988.

INSTANT TEST Page 1-21

Which of the following statements is/are nonsense? Why?

- (1) Albert was #1 out of the top 3 in the school ranking and Betty was #2, so Betty must have gotten two-thirds of the score that Albert got.
- (2) The police car was going 100 mph and my car was going 50 mph; thus the police car was going twice as fast as mine.
- (3) It's 80 miles to North Galanga from Balbuto and it's 160 miles to South Galanga from Balbuto. Therefore it is twice as far from Balbuto to South Galanga as it is from Balbuto to North Galanga.
- (4) Bathun's car gets 40 miles per gallon (mpg) of gasoline on average whereas Carlita's car gets 20 mpg. Therefore Bathun's car is twice as efficient (40/20) as Carlita's in terms of mpg.
- (5) Bathun's car takes 0.025 gallons per mile (gpm) of gasoline on average whereas Carlita's car takes 0.05 gpm. Therefore Bathun's car is half as expensive to run (0.025/0.05) as Carlita's in terms of gpm.
- (6) Davido's car takes 1st place in the WildCircles gasoline efficiency contest whereas Eduardo's car takes 10th place in the same contest. Therefore Eduardo's car is 1/10th as efficient as Davido's car.
- (7) I assigned the Furuncle automobile a fuel efficiency score of 4 out of 5 in importance for choosing a new car and maximum speed a score of 4 out of 5 in importance. The Furuncle model thus scored a total of $4 + 4 = 8$ out of 10 possible points. In contrast, the Putrescent model scored 3 out of 5 and a miserable 1 out of 5, respectively, on those two evaluations. Therefore the Putrescent model, with $3 + 1 = 4$ as its total score was half as good as the Furuncle model overall.
- (8) The Rigellian Robots spaceball team was 1st, 4th, 2nd, 1st, and 3rd in the last five Pan-Galactic Games. Therefore, the Rigellian Robots spaceball team has averaged $(1+4+2+1+3)/5 = 11/5 = 2.2$ nd place in the last five Pan-Galactic Games.
- (9) These Venusian pandas come in three colors: infrared (color 1), ultraviolet (color 2) and pink (color 3). In the sample under consideration, there are 5 infrareds, 2 ultraviolets and 4 pinks, thus giving an average color of $(1*5 + 2*2 + 3*4)/(5+2+4) = 21/11 = 1.9$.
- (10) Albert scored 800 on the math SAT and Bruce scored 400 on the same test. The SAT scores have a minimum of 200 as the lowest possible score. Thus Albert scored $(800/400) = 2$ times as many points as Bruce in the SAT.
- (11) Cathy scored 800 on the math SAT and Daniella scored 400 on the same test. The SAT scores have a minimum of 200 as the lowest possible score. Thus Cathy scored $(800-200)/(400-200) = 3$ times as many points as Daniella in the SAT.

2 Accuracy, Precision, Sources of Data, Representing Data

2.1 Accuracy, Precision, and Being Correct

In statistics and science in general, *accuracy* is the degree of closeness to reality or truth. Thus if a restaurant has sold exactly 244 pizzas today, the accurate number is 244; 243 and 245 are less accurate approximations or can be simply described as *inaccurate* or *wrong*.

Precision is the degree of uncertainty in a measure; in our restaurant example, 243, 244 and 245 are equally precise even though 243 and 245 are inaccurate.

Continuing our pizza example, if we decide to count the 244 pizzas in dozens, describing the sales as *20 ½ dozen pizzas* is perfectly accurate and perfectly precise. Describing the sales as *20.3333 pizzas* is less precise and slightly less accurate: one-third is actually an infinitely repeating decimal, so 0.3333 is a little less than 0.333333333.... Estimates of *20.333, 20.33, 20.3, 20* or “*more than 10 and less than 30*” dozen are all decreasingly precise and also decreasingly accurate because of the decreasing precision. However, if we imagine that there are 30 pizzas in an order, the values 30., 30.0, 30.00, 30.000 and so on may have increasing precision, but they all have equal accuracy.

A different issue is *being correct* in the *choice* of statistical methods and in the *execution* of those methods. Many applications of statistics provide significant benefits to users – if the analyses and calculations are correct. If the wrong statistical procedure is applied to a problem or if a valid procedure is performed sloppily, resulting in incorrect results, the consequences can be severe. To emphasize the importance of being careful, precise and correct, I have consistently given a zero grade to incorrect responses to statistical assignments regardless of the pathetic mewling of students used to getting partial credit for trying to apply the right technique.

In the real world of work, no one legitimately does statistical analysis just to look good. Presenting a mistaken analysis of which marketing campaign had the best results in selling a new product can result in collapse of a new company. Make a mistake in a critical calculation and you could be fired. But making a mistake in the calculations of how much of a new drug should be administered to patients as a function of their weight could make sick people even sicker – or kill them. Miscalculating the effects of temperature on an O-ring on a space shuttle booster actually did kill astronauts.^{24,25}

Students must get used to *checking their work* as professionals do. Make a habit of checking what you’ve done before you move on to the next step. When you are using computer programs that you have written or entering values into routines from statistical packages, you will find it helpful to try your calculations with data for which you know the answers.

²⁴ (Stathopoulos 2012)

²⁵ **Personal note:** When I was teaching applied statistics in Rwanda, Africa from 1976 through 1978, my students were astonished at being given zeros for having made arithmetic errors. However, almost all of the students in my classes in the Faculté des sciences économiques et sociales at the Université nationale du Rwanda were headed for government jobs, so I put the issue in terms that made sense to them. I told them to imagine that they were responsible for estimating how much of the international development aid should be spent on Hepatitis C vaccination programs in the struggling nation. They could base their estimates of appropriate costs on knowledge of the unit costs of the vaccine, the costs of training and paying health workers, the number of people to be vaccinated, and the statistics on average losses of the vaccines through accident and errors. So what would happen if they made an arithmetic mistake in their calculations and failed to notice that they were, say, underestimating the appropriate cost by an order of magnitude? Imagine that instead of allocating the equivalent of US\$100,000 to the project, they erroneously reserved only US\$10,000 – and 790 men, women and children succumbed to Hepatitis C? Would they argue that they should get partial credit despite the mistake in their calculations? Would the victims’ families agree? There was silence in the classroom that day, and no one ever complained again about having to check their work step by step as they went through their problem solving.

2.2 Significant Figures

Significant figures reflect the degree of uncertainty in a measurement.²⁶ If we count 29 Ping-Pong balls in a bag, the number 29 has 2 significant figures and is exact: there is no uncertainty about it. However, if we weigh the 29 Ping-Pong balls and have a total weight of 115.2 g, the weight has four significant figures and represents a value that could actually be between 115.150 g and 115.249 g. This weight to four significant figures is expressed in *scientific notation* as 1.152×10^2 g or as 1.152e2 g.

Here are some examples of constants that can be expressed with different numbers of significant figures:

- π represents the ratio of the circumference of a circle to its diameter. It can be expressed as, say, 3.14159 using six significant figures but has been calculated to more than a million decimal digits;²⁷
- e , the limit of $(1 + 1/n)^n$ and the base of natural logarithms, to nine significant figures is 2.71828183 but continues without limit to the number of digits that can theoretically be computed;²⁸
- c , the speed of light in a vacuum and the maximum speed of anything in the known universe except science-fiction ships using faster-than-light travel, can be given as 299,792,458 meters per second (a number with nine significant figures also); if it is expressed as 186,282 miles per second then it has six significant figures; and if it is expressed as 1.079e9 km/hour then it has four significant figures;²⁹
- *Avogadro's number* to seven significant figures is 6.022142×10^{23} particles / mole (with many more significant digits available – but not an infinite number of significant figures because molecular weight involves discrete particles and therefore can in theory be counted to an exact integer). This number is the number of atoms in 12 grams of pure carbon-12 and defines a *mole* of a material.³⁰

Most other numbers we work with are naturally variables: salaries, returns on investment, quality percentages, customer satisfaction responses, and contaminant measurements. So to how many significant figures should we express them?

The issue depends on how variable a measure is compared with its size in the scale of measurement we want to use. For example, suppose we are looking at the number of Internet Protocol (IP) packets arriving every second at a specific port in a firewall; this number fluctuates from moment to moment, but we can actually know exactly how many packets there are from the log files kept by the firewall software. However, if someone asks, “How many packets per second were reaching port 25 during the last hour?” answering with a list of 3,600 numbers – one total per second for each of the 3,600 seconds of the hour – is likely to be met with astonishment, if not hostility. We naturally want to express the number using a summary, so we compute an average (we will study those in detail later) and answer using the average.

INSTANT TEST

Suppose a cubic nanometer of pure Ultronium is estimated to have 17,298 atoms – which has 5 significant figures.

Express this number in scientific notation with the e format (e.g., 542.3 = 5.423e2) to 1 significant figure, then to 2 significant figures, then to 3 significant figures and finally to 4 significant figures.

Notice also what happens to 17,298 when you express it to 4 and then to 3 significant figures. . . .

²⁶ (Morgan 2010)

²⁷ (University of Exeter 2012)

²⁸ (Wolfram Mathworld 2012)

²⁹ (Fowler 2009)

³⁰ (Cambell 2011)

2.3 Determining Suitable Precision for Statistics

How precisely should we express derived statistics such as an average? Consider the following two cases:

- Case 1: there were a total of 129,357,389 packets received over the hour for an average of exactly 49,752.8419230769 packets per second. The lowest rate was 7,288 per second and the highest rate was 92,223 per second.
- Case 2: the same number of packets was received, so the average was the same as in Case 1, but this time the lowest rate was 49,708 packets in a second and the highest rate was 49,793 packets in a single second.

So how should we express the data and the averages in these two cases?

Precision refers to the “closeness of repeated measurements of the same quantity” as Robert R. Sokal and F. James Rohlf put it in their classic statistics textbook³¹ which influenced generations of biologists.^{32,33,34}

In both of our examples, the 15 significant figures of 49,752.8419230769 seem pointlessly and incorrectly precise. That number implies that the real average value is approximately between 49,752.84192307685 and 49,752.84192307695 but such precision is misleading when the actual observed variation is between 7,288 and 92,223 packets per second in the first case and 49,708 and 49,793 packets per second in the second case.

A general approach for deciding on appropriate precision is that we should express results for original observations so that the *range* (difference between the smallest and the largest observed values) is *divided into roughly 30 to 300 steps or divisions*. In addition, *derived statistics* based on our data typically have *one more significant digit* than the original data.

The easiest approach to deciding on significant figures is to estimate the number of steps at different levels of precision. After a while, one becomes good at guessing a reasonable degree of precision. Using our two cases,

- Case 1: a range from a minimum of 7,288 per second up to a maximum of 92,223 per second means that
 - One significant figure would result in a value of 90,000 packets per second (i.e., 9 units of 10,000 packets) for the upper limit and thus the lower limit would be 10,000 packets per second (1 unit of 10,000 packets – not 7,000, which would imply a precision of thousands rather than of tens of thousands).
 - Reporting to *two* significant figures would give us values of 92,000 and 7,000; the number of thousands would be $92 - 7 + 1 = 86$ steps or divisions (thousands).³⁵ That would be perfect: right in the middle of the 30-300 steps rule. Notice that the minimum has *fewer significant figures* in this case to maintain consistency with the upper value.
 - To estimate how many steps there would be if you increased the number of significant figures by one (e.g., 92,200 in which we tally the number of hundreds), you can either multiply the previous number of divisions at the current level (86) by 10 for a rough approximation (~ 860 divisions) or you can insist on being precise and do the arithmetic (922 hundreds – 73 hundreds + 1 = 850 divisions) to see how many steps there would be with *three* significant figures. That's too many divisions for the raw data.
 - However, the average *would* be reported with one more significant figure than the two used for the data, thus giving *three* significant figures for derived statistics. Suppose the average were, say, 49,758.268. This average would thus be reported reasonably with three significant

³¹ (Sokal and Rohlf, Biometry: The Principles and Practice of Statistics in Biological Research 1981)

³² I studied the First Edition in 1969 and helped proofread the Second Edition in 1980, getting mentioned in the acknowledgements [beams proudly] and also receiving a surprise check for U\$400 (about U\$1,200 in 2012 dollars) that enabled me to buy an excellent touring bicycle!

³³ (I'm really old.)

³⁴ (Compared to you).

³⁵ Why is there a +1? Because if you subtract a lower number from a higher number, you don't get the total number of values including them; e.g., $6 - 3 = 3$ but there are actually 4 values included in that range: 3, 4, 5 & 6.

figures as 49,800 packets per second. Note that there cannot be a decimal point there because it would incorrectly suggest that we were reporting using five significant figures.

- Case 2: the lowest rate was 49,708 packets in a second and the highest rate was 49,793. Let's also suppose that the average was as above: 49,758.268.
 - Using what we found for Case 1 (*three* significant figures) as a guess, we see that the range based on the minimum and maximum would be reported as $49,700 - 49,800 =$ only *two* steps.
 - It's easy to see that if we use four significant figures, there would be $49,790 - 49,710 + 1 = 81$ steps which is perfectly within the 30-300 guidelines.
 - So four significant figures would be great for these raw data and therefore the average would be reported with one more significant figure = 5. The average (e.g., 49,758.268) would thus be reported reasonably as 49,753 packets per second. It would be OK to use a decimal point (49,753.) If such a number ended in a zero (e.g., some other average computed as, say, 49,750) it would be reported with a terminal decimal point (47,750.) to be clear that it was expressed with 5 significant figures.

To repeat this last note on significant figures of numbers ending in zero that are expressed in ordinary notation: be careful about the use of a decimal point when you have relatively few significant figures. For example, if a number expressed to 3 significant figures is 12,300 it would be a mistake to write it as 12,300. (note the period) because that would imply that we were reporting to 5 significant figures. This kind of nuisance explains why we often choose to use scientific notation when correct communication of numerical values is paramount.

Scientific notation uses an integer between 1 and 9 and is followed by an appropriate number of decimal digits. For example,

- 1,234.56789 expressed to 6 significant figures would be shown as 1.23457×10^3 or as 1.23457e3 (this latter format is particularly useful because it works in Excel)
- The same number reduced to 4 significant figures would be shown as 1.235×10^3 or 1.235e3
- With only significant figures, the number would be shown as 1.23×10^3 or 1.23e3
- For numbers smaller than 1, count the number of places you have to move the decimal point to arrive at the integer portion of the scientific notation. Thus a value of
 - 0.0001234 to 3 significant figures would be 1.23×10^{-4} or 1.23e-4
 - 0.00456789 to 5 significant figures would be 4.5679×10^{-3} or 4.5679e-3

INSTANT TEST Page 2-5

- (1) The average cost of shares for Urgonium Corporation over the last month is calculated as 123.456789 credits. Express the average in non-scientific notation using
- 2 significant figures
 - 3 significant figures
 - 4 significant figures
 - 8 significant figures
- (2) Express the average above in scientific notation using
- 2 significant figures
 - 3 significant figures
 - 4 significant figures
 - 8 significant figures

(cont'd)

- (3) The smallest cost observed (the minimum) for Urgonium Corporation shares was 118 credits and the largest value observed (the maximum) was 128 credits. How many steps would there be in the range if you used 2, 3, 4, or 8 significant figures for these limits?
- (4) The minimum number of retaining bolts per strut on the Garagano Narrows Bridge is 17,826 and the maximum number is 22,012. The average number is 19,943.24 Which of the following is an appropriate number of significant figures for these data? Why?
- a) 17,826.0 | 19,943.24 | 22,012.0
 - b) 17,826. | 19,943.2 | 22,012.
 - c) 17,830. | 19,943. | 22,010
 - d) 17,800 | 19,940 | 22,000
 - e) 18,000 | 19,900 | 20,000
 - f) 20,000 | 20,000 | 20,000

2.4 Sources of Real Statistical Data

At many points in this course, you will be given the opportunity to apply newly learned methods to real-world data. There are many sources of such data; e.g., looking at United States resources first,

- US Bureau of Labor Statistics³⁶ has a wealth of economic data such as
 - Inflation & Prices
 - Unemployment
 - Employment
 - Spending & Time Use
 - Pay & Benefits
 - Productivity
 - Workplace injuries
 - International economic comparisons
- US Bureau of Justice Statistics³⁷ provides data about crime and justice such as
 - Corrections
 - Courts & Sentencing
 - Crime Type
 - Criminal Justice Data Improvement Program
 - Employment & Expenditure
 - Federal
 - Law Enforcement
 - Victims
- US Census Bureau³⁸ provides information about
 - People & Households Business & Industry Geography
 - Fraudulent Activity & Scams
 - Census Bureau Data & Emergency Preparedness
 - The Statistical Abstract of the United States, “the authoritative and comprehensive summary of statistics on the social, political, and economic organization of the United States. Use the Abstract as a convenient volume for statistical reference, and as a guide to sources of more information both in print and on the Web.”
- US Bureau of Transportation Statistics³⁹ has the “mission is to create, manage, and share transportation statistical knowledge with public and private transportation communities and the Nation.”
- US Centers for Disease Control and Prevention (CDC)⁴⁰ provide verified data about health issues such as
 - Diseases & Conditions
 - Emergency Preparedness & Response
 - Environmental Health
 - Life States & Populations

³⁶ < <http://www.bls.gov/> >

³⁷ < <http://bjp.ojp.usdoj.gov/> >

³⁸ < <http://www.census.gov/> >

³⁹ < <http://www.bts.gov/> >

⁴⁰ < <http://www.cdc.gov/> >

- Healthy Living
- Injury, Violence & Safety
- Travelers' Health
- Workplace Safety & Health
- US Washington Headquarters Services (WHS)⁴¹ whose Pentagon Library offers links to many sources of statistical information, especially focused on military matters.
- World Health Organization (WHO)⁴² of the United Nations provides extensive information (in many languages) about health and wellbeing around the world such as
 - Mortality & Health Status
 - Diseases
 - Coverage of health services
 - Risk factors (alcohol, nutrition, overweight & obesity, tobacco)
 - Health Systems
- Pew Research Center covers a vast array of topics⁴³ comprising
 - Demography
 - Domestic Policy
 - Economics
 - Election '08
 - Election '12
 - Energy and Environment
 - Global Attitudes/Foreign Affairs
 - Immigration
 - Internet and Technology
 - Legal
 - News Media
 - Politics and Elections
 - Polling History
 - Public Opinion
 - Religion
 - Research Methodology
 - Social Trends
- Sports statistics sites⁴⁴
- Internet World Stats⁴⁵ provides global information and details broken down by region
 - Africa
 - America
 - Asia
 - Europe

⁴¹ < <http://www.whs.mil/> >

⁴² < <http://www.who.int/research/en/> >

⁴³ < <http://pewresearch.org/topics/> >

⁴⁴ Use a search engine with the search term *sports statistics* to find many Websites providing such information

⁴⁵ < <http://www.internetworldstats.com/stats.htm> >

- European Union
- Mid-East
- Oceania
- Scientific and Professional Journals
 - Visit your university or local library to enquire about how to access to thousands of scientific and professional journals available online that often include detailed statistical information
 - At Norwich University, the Kreitzberg Library Databases⁴⁶ are available online for all members of the Norwich community and for all disciplines.

In general, most government and survey sources provide direct access to downloadable data in various formats. If you need to work with the original data used in a scientific or professional publication, you may be able to ask the analysts for access to raw data files; although you may not always gain access, it's worth trying. You should explain exactly what you intend to do with the data and should offer to provide the original analysts with the results of your work. If the data are not in the public domain (e.g., published by the US government) then you must explicitly ask for permission to use copyright material in your work.

Failing direct access to usable data formats, you can use conversion programs; for example, Adobe Acrobat Professional includes tools for recognizing and interpreting text from scanned documents or other PDF files. However, optical character recognition (OCR) of all types is never perfect; you should check the conversions with minute attention to detail. Publishing an analysis that uses incorrect values is a career-killer.

Finally, you may simply have to re-enter data; again, be sure to double-check the accuracy of the transcription. Remember, in addition to ruining your reputation, appropriating data without permission or misstating the data from other people's work could lead to lawsuits for copyright infringement and for defamation (putting someone in false light). Misappropriation without precise identification of the origin of the data (or text) you use may also result in punishment for *plagiarism*.

INSTANT TEST P 2-8

Using your own particular interests, look up some interesting statistical information from among the resources listed or others and report on what you find by posting a note in the appropriate discussion group on NUoodle.

Examples of potentially interesting statistics:

- Examine the changing distributions of wealth and income in the USA over the last several decades
- Compile information about educational attainment in different demographic sectors
- Evaluate the US position on health care delivery and population health compared with other countries' records
- Gather information about construction trends – what kinds of building materials are being used – and how have these patterns changed over time?
- Evaluate fuel efficiency statistics for different kinds of vehicles; e.g., various types of cars, trains, ships and airplanes
- Track the popularity of different sports or teams over time and in different regions of the USA or of the world
- Learn about the changing patterns of Internet, cell phone, computer usage

⁴⁶ < <http://www.norwich.edu/academics/library/databases.html> >

2.5 Representing Data

Both to clarify our own understanding of data and to help others, we often show data in various forms – tables, charts, graphs and other formats. The most important principle in choosing a data representation is clarity: the format must communicate our intention with a minimum of ambiguity.

2.6 Presenting Raw Data

There are many situations in which the specific sequence of data collection is valuable for our research – either because it reflects a time sequence relating the observations to the external world (for example, stock-market price fluctuations that may be related to news events).

There may also be interest in internal data sequences such as the likelihood that a particular pattern will follow a specific sequence of data (this kind of analysis includes what is called *Markov chain* analysis) such as machine learning algorithms to identify customer buying patterns (more likely to buy a certain brand if the customer has just bought two other products in a row) or network attack patterns (e.g., perhaps probes on ports 80 and 25 in the last five minutes may increase the likelihood of probes on port 20 within the next 10 minutes).

Although we rarely publish thousands of individual observations in a table, it is frequently the case that we must produce the original data for inspection. A simple approach is to list the observations in a single column or row, but with thousands of values this method is impractical for publication. **Error! Reference source not found.** shows the top and bottom parts of a simple table of 50 rows (a label plus 49 values).

Creating multi-column tables from data in a single column is not trivial in EXCEL; there are no facilities in that program for automatically showing parts of the single column in side-by-side segments. However, word-processing packages such as MS-WORD make it easy to convert raw data into a useful and professionally presentable format. These packages include facilities for creating columns, so it becomes trivially easy to convert the data.

We can import the data into a column and then format them as text into several columns, as shown in Figure 2-2. Be careful not to make the columns so narrow that numbers begin breaking into different lines. Reduce the number of columns if that happens.

Figure 2-1. Sample of raw data in Excel.

A
1 Raw Data
2 282
3 211
4 433
5 229
6 126
7 254
8 334
9 373
10 462
11 170
48 241
49 428
50 151

Figure 2-2. Simple presentation of data in columns.

Raw Data:

282	345	428	224	141
211	472	176	295	127
433	351	148	358	233
229	269	159	162	322
126	492	471	139	306
254	192	390	115	263
334	460	303	352	241
373	419	374	223	428
462	287	161	495	151
170	222	452	260	

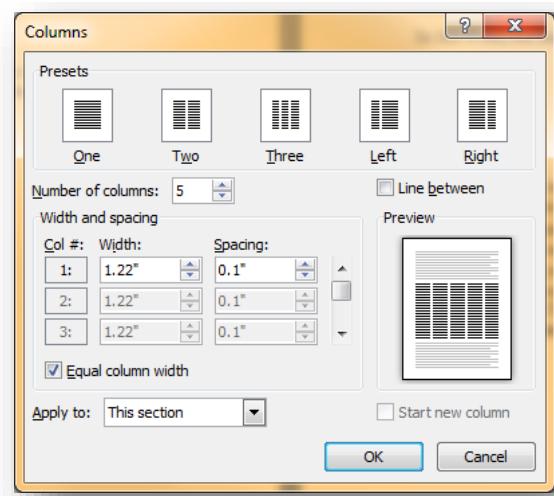
In MS-WORD, the dialog for **Columns** offers many options, as shown in Figure 2-3. It's easy to determine the number of columns and the spacing between each pair. There's also an option for adding a vertical line between columns.

Enhancing a multi-column table of data is easy in WORD. For example, Figure 2-4 shows a few rows of data in an EXCEL table.

Figure 2-3. MS-WORD dialog box for defining columns.

Figure 2-4. Multi-column table in Excel.

	A	B	C
1	Variable 1	Variable 2	Variable 3
2	282	6.5	8.0E+03
3	211	9.4	3.4E+04
4	433	9.0	3.2E+04
5	229	6.6	4.3E+04
6	126	7.9	2.4E+04
7	254	5.3	1.3E+04
8	334	6.6	3.4E+04
9	373	9.1	2.4E+04
10	462	8.9	1.2E+04
11	170	6.2	1.7E+04



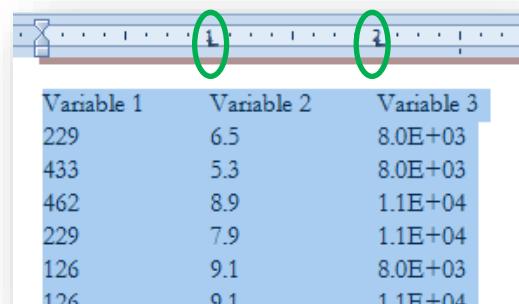
Once the data are pasted into WORD, columns are separated using the **Tab** character. In Figure 2-6, the data are shown using the **Show Paragraph** function (**¶**), which is a toggle that is also accessible using the key-combination **Shift-Ctrl-***. The arrows represent the **Tab** character and the paragraph symbols show the end-of-line character. These characters are not normally visible – they are in the snapshot to show how columns are converted from EXCEL into WORD –columns are separated by **Tab** characters in every row.

The simple approach to aligning the labels with the columns is to highlight the data and insert appropriately seven spaced **Tab** marks in the WORD ruler, as shown in Figure 2-5. We'll come back to cosmetic enhancements of tables in WORD in a few pages.

Figure 2-5. Data pasted into WORD showing tab and paragraph marks.

```
Variable 1    Variable 2    Variable 3¶
282    6.5    8.0E+03¶
211    9.4    3.4E+04¶
433    9.0    3.2E+04¶
229    6.6    4.3E+04¶
126    7.9    2.4E+04¶
254    5.3    1.3E+04¶
334    6.6    3.4E+04¶
373    9.1    2.4E+04¶
462    8.9    1.2E+04¶
170    6.2    1.7E+04¶
```

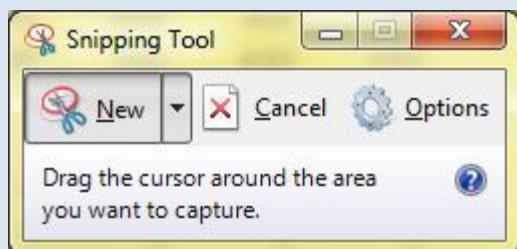
Figure 2-6. Tab marks in ruler used to align labels with imported data.



INSTANT TEST P 2-11

Using data that you find on the Web to reflect your own interests, practice the following skills:

- Capturing a screenshot of an interesting table (e.g., using Windows 7 "Snipping Tool" or clicking on a window and pressing Alt-PrtScn)



- Pasting an image into a WORD document using Ctrl-Alt-V
- Copying data from a Web page by highlighting them and then pasting them into a WORD document in various formats (Ctrl-Alt-V again)
- Taking the same data from the Web page you found and pasting them into an Excel spreadsheet
- Copying a table created in Word and pasting the data into Excel in various formats
- Copying a table created in Excel and pasting the data into Word in various formats
- Writing out a list of numbers with spaces between them and pasting them into Excel (not pretty, eh?)
- Converting the data in your Word file into the same list with TAB characters between them instead of spaces and then pasting the new list into Excel
- Creating a list of e-mail addresses in Excel, then copying the list into Word and replacing the TAB characters by the string <;^P> (not including < and >) to create a semicolon-delimited list with a space after every semicolon so you can paste the list into an e-mail BCC field.

3 Sorting, Backups and Enhanced Tables

3.1 Sorted Lists

There are many cases when we find a sorted list helpful. WORD has simple sort functions, as does EXCEL. Such lists help the viewer grasp the range of variation better than an unsorted list and immediately reveal the minimum and the maximum values.

3.2 Simple Sorting in WORD

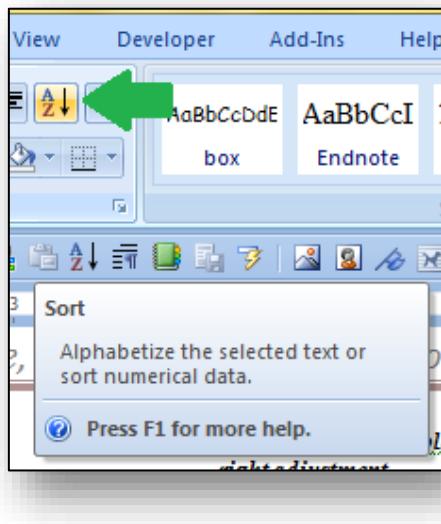
Figure 3-1 shows data that the user wants to sort so that each column will be sorted in ascending order – all the data in the first column in order, with any values that are the same in the first column sorted by the value of the second column, and then again using the third column.

Figure 3-1. Table of values to sort in WORD.

Variable 1	Variable 2	Variable 3
229	6.5	8.0E+03
433	5.3	8.0E+03
462	8.9	1.1E+04
229	7.9	1.1E+04
126	9.1	8.0E+03
126	9.1	1.1E+04
492	7.4	1.1E+04
192	7.4	1.1E+04
126	5.3	1.1E+04
462	6.6	1.1E+04
211	9.4	1.3E+04
126	7.9	1.1E+04
433	9.0	1.1E+04
170	7.1	1.1E+04
460	8.6	1.1E+04
229	6.6	1.1E+04
229	7.9	8.0E+03
211	9.4	8.0E+03
428	6.6	1.3E+04
211	9.5	1.1E+04

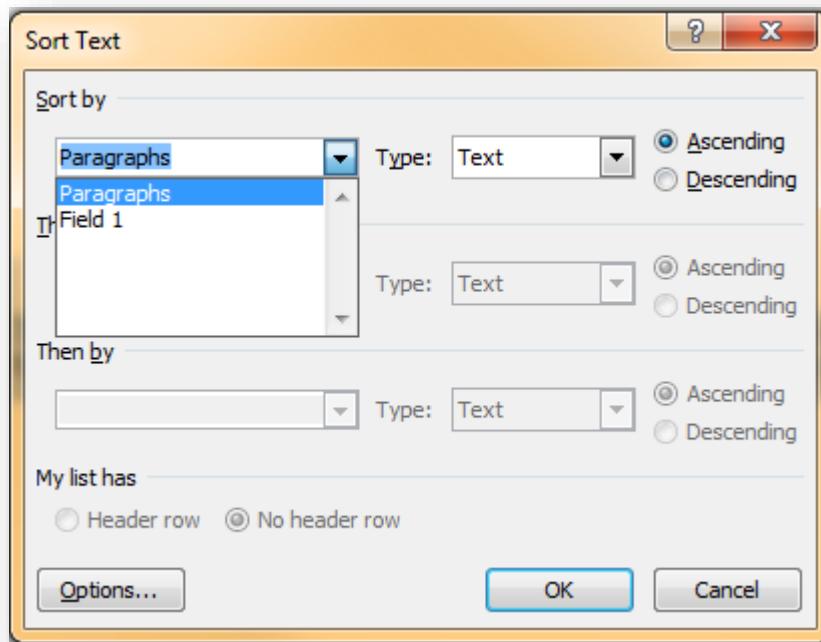
In WORD 2007, highlight the text to sort and select the **Sort** button (Figure 3-2)

Figure 3-2. Word sort button.



to start the **Sort Text** dialog box (Figure 3-3).

Figure 3-3. Word sort dialog box showing options of what to sort by.



The sort can be ascending (e.g., from A to Z) or descending (Z to A). Clicking on the **Header row** buttons tells WORD whether to include the top row (if it's a header) or include it in the sort (if it's not a header). Selecting paragraphs is useful for blocks of information such as multi-line text in bullet points. Selecting fields uses the Tab characters we discussed earlier as separators defining Field 1, Field 2 and Field 3 corresponding to columns. As you can see, the sort instructions in Figure 3-5 specify fields rather than paragraphs and the check in **Header row** converts the column references to the names in the first row.

Stipulating **Number** as the **Type** of variable is important, because sorting by text or date can result in different sort orders. Figure 3-4 shows the pull-down menu for the **Type**.

Figure 3-5. Sort instructions in Word for three columns of numerical information.

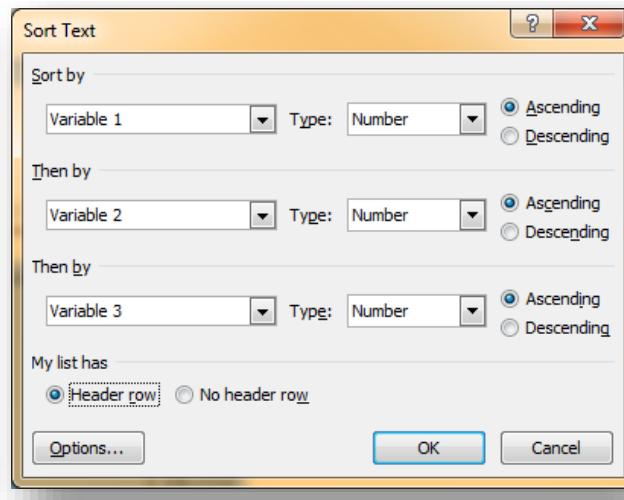
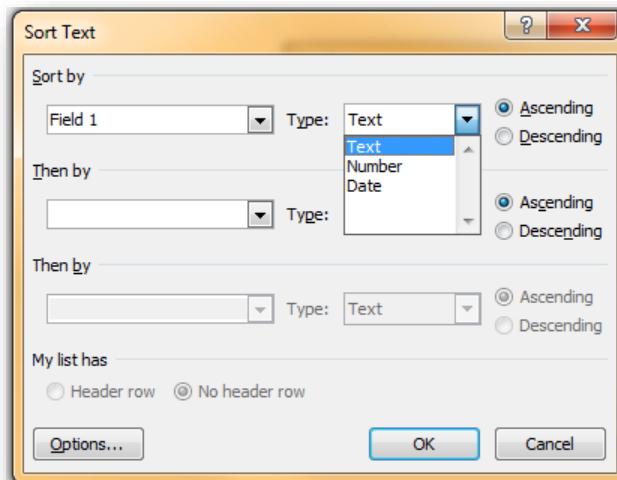


Figure 3-4. Word sort options showing choice of three types of sort to use.



As shown in Figure 3-6, the sort results in sequences where values of the first column are in ascending order. Within groups of identical values of Variable 1 (e.g., 126), the lines (also called *records* from the terminology of databases) are sorted according to the value of Variable 2, the second column. For Variable 1 = 126, the records where Variable 2 is the same (= 9.1) are together and sorted by the value in the third column, Variable 3 (first 8.0E+03 then 1.1E+04).

Figure 3-6. Result of Word sort instructions.

Variable 1	Variable 2	Variable 3
126	5.3	1.1E+04
126	7.9	1.1E+04
126	9.1	8.0E+03
126	9.1	1.1E+04
170	7.1	1.1E+04
192	7.4	1.1E+04
211	9.4	8.0E+03
211	9.4	1.3E+04
211	9.5	1.1E+04
229	6.5	8.0E+03
229	6.6	1.1E+04
229	7.9	8.0E+03
229	7.9	1.1E+04
428	6.6	1.3E+04
433	5.3	8.0E+03
433	9.0	1.1E+04
460	8.6	1.1E+04
462	6.6	1.1E+04
462	8.9	1.1E+04
492	7.4	1.1E+04

3.3 Simple Sorting in EXCEL

The EXCEL Sort & Filter menu shown in Figure 3-7 offers both simple sorts and more complex sorting options (the Custom Sort... option). Filter is a useful function that shows selected values; we'll discuss that function later.

The easiest form of sorting simply sorts by the first column in a selected area of the spreadsheet. For example, Figure 3-8 shows information about sales for various salespeople. It would be convenient (and normal) to show these data sorted by the name of the salesperson. By highlighting the entire table and then using the Sort AZ↓ button, it's easy to sort automatically using that first column. The results are shown in Figure 3-9.

Figure 3-8. Excel table with unsorted sales figures.

Salesperson	Sales
Marcie	\$ 909,460
Darlene	\$ 1,153,452
George	\$ 1,389,164
Charlene	\$ 630,871
Bob	\$ 979,272
Harold	\$ 398,478
Louise	\$ 1,193,042
Holly	\$ 1,020,585
John	\$ 1,159,705
Albert	\$ 1,644,140
Samuel	\$ 1,704,159
Mike	\$ 324,598
Robert	\$ 1,305,230
Edward	\$ 516,562
Francine	\$ 844,769
Sally	\$ 951,054
Victor	\$ 826,891
David	\$ 130,049

Figure 3-7. Excel sort & filter menu.

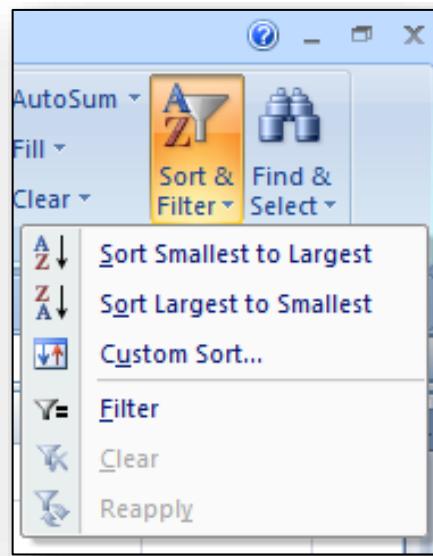


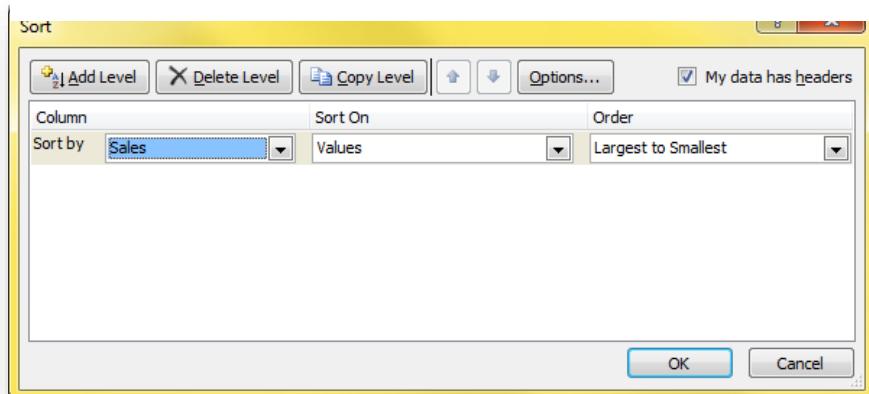
Figure 3-9. Sales data sorted automatically by first column.

Salesperson	Sales
Albert	\$ 1,644,140
Bob	\$ 979,272
Charlene	\$ 630,871
Darlene	\$ 1,153,452
David	\$ 130,049
Edward	\$ 516,562
Francine	\$ 844,769
George	\$ 1,389,164
Harold	\$ 398,478
Holly	\$ 1,020,585
John	\$ 1,159,705
Louise	\$ 1,193,042
Marcie	\$ 909,460
Mike	\$ 324,598
Robert	\$ 1,305,230
Sally	\$ 951,054
Samuel	\$ 1,704,159
Victor	\$ 826,891

3.4 Advanced Sorting in EXCEL

How do we sort by a column that is not the leftmost in our range? To do so, activate the **Sort & Filter** functions shown in Figure 3-7 and click on the **Custom Sort...** option. The menu shown in Figure 3-10 appears and one can select a sort field (in this example, the Sales figure column) and the direction of the sort (here, a downward sort putting the highest sales figures at the top of the sorted results). It's useful to select the label row as well as all the data and to click the **My data has headers** box: the names defined in the label row then appear in the dialog box instead of just showing the EXCEL column name (Column A, Column B, and so on). The rest of the process is similar to what you have learned about advanced (multi-field) sorting in WORD.

Figure 3-10. Excel sort options showing selection of second column (Sales) as the primary sort field.



The results of this sort are shown in Figure 3-11.

Figure 3-11. Data sorted by sales with highest sales at top.

Salesperson	Sales
Samuel	\$ 1,704,159
Albert	\$ 1,644,140
George	\$ 1,389,164
Robert	\$ 1,305,230
Louise	\$ 1,193,042
John	\$ 1,159,705
Darlene	\$ 1,153,452
Holly	\$ 1,020,585
Bob	\$ 979,272
Sally	\$ 951,054
Marcie	\$ 909,460
Francine	\$ 844,769
Victor	\$ 826,891
Charlene	\$ 630,871
Edward	\$ 516,562
Harold	\$ 398,478
Mike	\$ 324,598
David	\$ 130,049

INSTANT TEST Page 3-6

Using data you find in a publication of interest, enter at least ten rows of data with at least three columns into a spreadsheet.

- (1) Sort the entire table of data alphabetically by the leftmost column in ascending order and then in descending order.
- (2) Practice sorting on other columns individually
- (3) Practice sorting on two columns, then on three columns.

3.5 Mistakes in Sorting

When working with these functions, highlighting all the data is essential. Leaving out a column from the sorted area and then sorting destroys the integrity of the data by mixing up parts of the observations.

For example, the European Spreadsheet Risks Interest Group (EuSpRiG) posted these reports from 2005 and 2006:

Aspiring Police Officers Say Exam Scores Were Botched (NBC13.com 8 Sep 2005)

“Some aspiring police officers who took a government exam said they were told they passed a big test, but found out later that they had actually failed. A national company called AON administered the test and told the board someone incorrectly sorted the results on a spreadsheet, so the names and scores were mismatched”, NBC 13’s Kathy Times reported.

“When they appealed, we went back to AON and asked them to check their scores, and when they audited, they discovered they made an error,” said Bruce Nichols, of the Jefferson County Personnel Board. Nichols has resigned.

Price mixup mars opening of lot sales Venice, Florida, Jan 1, 2006

A snafu in the posting to the web site of minimum bid prices for the first phase of North Port’s abandoned lot auction led to confusion as the cost of some lots seemingly tripled overnight. The appraiser hired by the county put the auction lot number, the property ID number and the minimum bid amount onto a spreadsheet in sequential order and, inadvertently, he did not sort the value column.⁴⁷

INSTANT TEST Page 3-7

Using the multi-column data you created for the Instant Test on the previous page,

- (1) Take a screenshot of your table and save the image as “properly_sorted” in whatever picture format you like (JPG, GIF, PNG....)
- (2) Now deliberately highlight all but the rightmost column and sort using the leftmost column. Take a screenshot of that table and save it as “badly_sorted”.
- (3) Study the effect of the bad sort so that you remember the consequences of this kind of error and never do it again!

⁴⁷ (EuSpRIG – European Spreadsheet Risks Interest Group 2012)

3.6 Making Backups of Your Work

As a routine precaution, EXCEL versions make AutoRecover backups every few minutes. If EXCEL crashes, you can restart and the latest autorecovery file will be opened and provide the latest status of the workbook.

It is also advisable to save the worksheet yourself before applying the sort for quick recovery after a serious error.

A good practice is to use version numbers (e.g., `filename_v04.xlsx`) and to increment the version number when starting to work again the next day on the file or before performing potentially dangerous operations.

For example, on starting to work in the morning with file `filename_v04.xlsx`, one could OPEN `filename_v04.xlsx` and immediately SAVE AS `filename_v05.xlsx`.

When you back up your work to an external medium (flash drive, USB or Firewire disk drive), you won't overwrite your only copy. For example, if you were to save your work as `filename.xlsx` before stopping work every day, your Tuesday evening backup would overwrite your Monday backup. If you made a terrible mistake on Tuesday, such as erasing an entire worksheet in your workbook, you would not be able to recover the deleted data. With `filename_v04.xlsx` on your external medium for Monday night and `filename_v05.xlsx` on Tuesday night, you could go back to `filename_v04.xlsx` and copy the data to paste into your damaged file.

Figure 3-12 below shows the first few lines of the listing of my own backup files for an earlier version of this textbook on a 1 TB external USB3 disk drive that I update from the C: drive every evening.⁴⁸

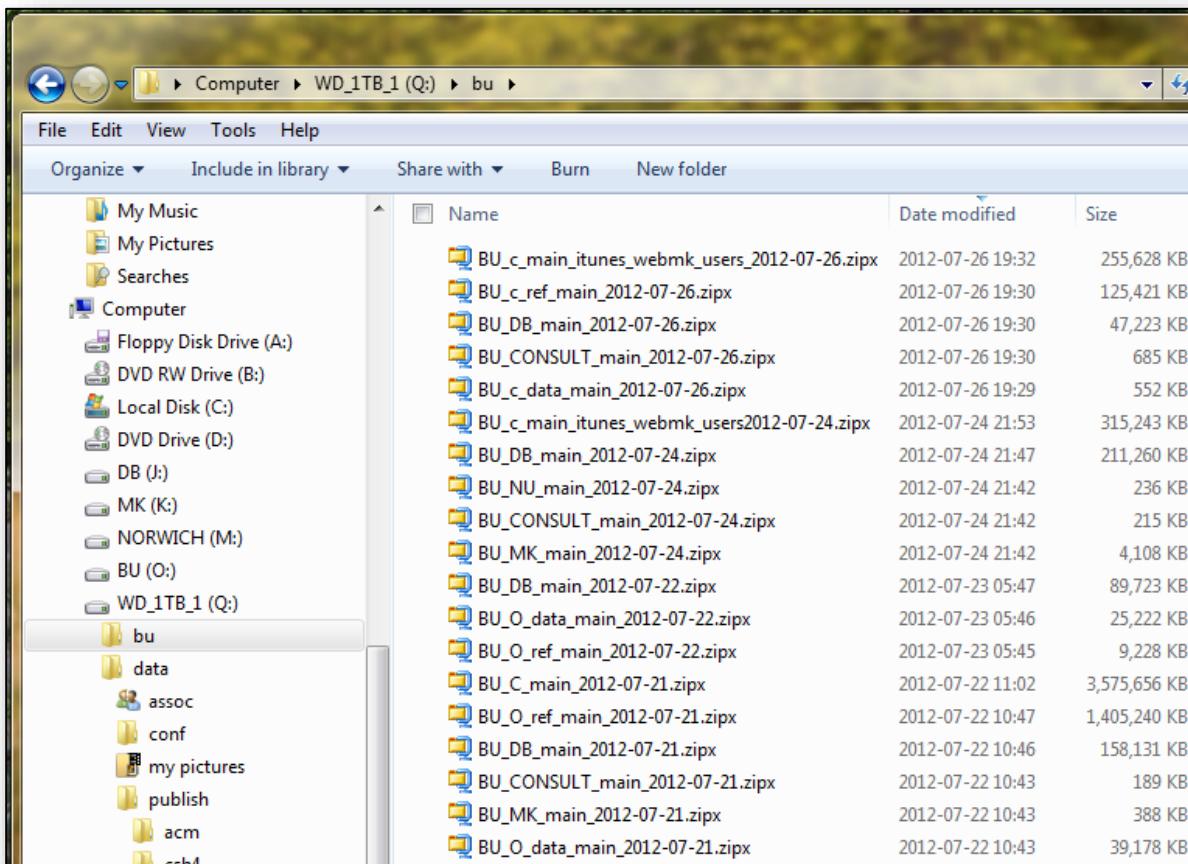
Figure 3-12. Backup files with version numbers for statistics textbook.

 Backup of statistics_text.wbk	2019-04-04 20:23
 qm213_WEB_shortcut	2019-04-04 20:33
 statistics_text.docx	2019-04-04 20:32
 statistics_text_v0-23-2.docx	2018-03-19 07:52
 statistics_text_v0-24-1.docx	2018-03-19 08:46
 statistics_text_v0-25-1.docx	2019-02-08 09:40
 statistics_text_v0-25-2.docx	2019-04-04 20:32

⁴⁸ I tend to make more frequent version backups than once a day because I detest having to redo work if I have an accident.

For completeness, here's a snapshot from some years ago of some general backup files for my main tower system on the same external disk. The ZIPX files are generated using WinZip,⁴⁹ a popular and inexpensive file compression utility that has features such as jobs that can create multiple separate backups with a single click.⁵⁰

Figure 3-13. General backups on external disk drive.



INSTANT TEST P 3-9

Practice creating versions of a file.

Open an Excel file and Save As < demo_v1.xlsx > and close that file. Open < demo_v1.xlsx > and make any change (add a random letter or number somewhere on a sheet). Save As < demo_v2.xlsx > and exit. Sort your file list by date with the newest files at the top. Look at the time stamp of the two versions you just created. Open the latest file and save it with an incremented version number. Remember this process when you work on your homework. And don't expect much sympathy from your instructor if you claim that the computer ate your homework files.

⁴⁹ <http://www.winzip.com/win/en/index.htm>

⁵⁰ For more guidance on how to use folders and backups, see "Organizing and Safeguarding Information on Disk." <http://www.mekabay.com/methodology/osiod.pdf>

3.7 Enhancing the Presentation of Tables

Both WORD and EXCEL offer extensive options for creating and formatting tables. Such tables are used in business, engineering, and science to summarize information; enhancing the tables using borders, colors, and title rows helps the reader understand the information quickly and correctly.

3.8 WORD Table Tools

WORD offers easy methods for creating small, simple tables without calculations. WORD 2007, for example, has a drop-down menu in the **Insert** tab that allows one to create a blank table form of a particular number of rows and columns using one's cursor to highlight the right shape and then lets one fill in data and format the table (see Figure 3-15).

Figure 3-15. Creating a new blank table in Word.

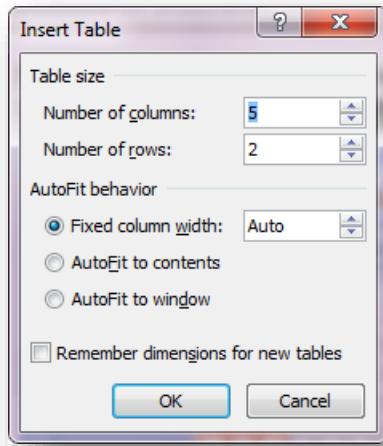
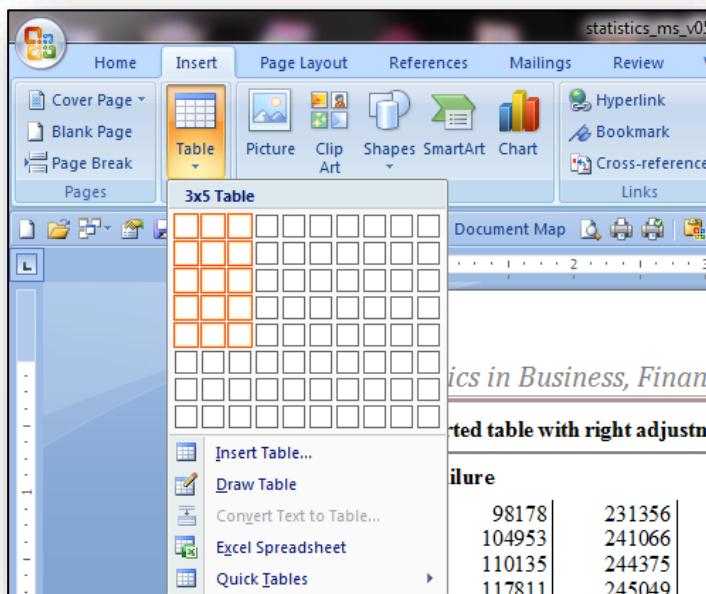


Figure 3-14. Insert Table pop-up menu in Word.



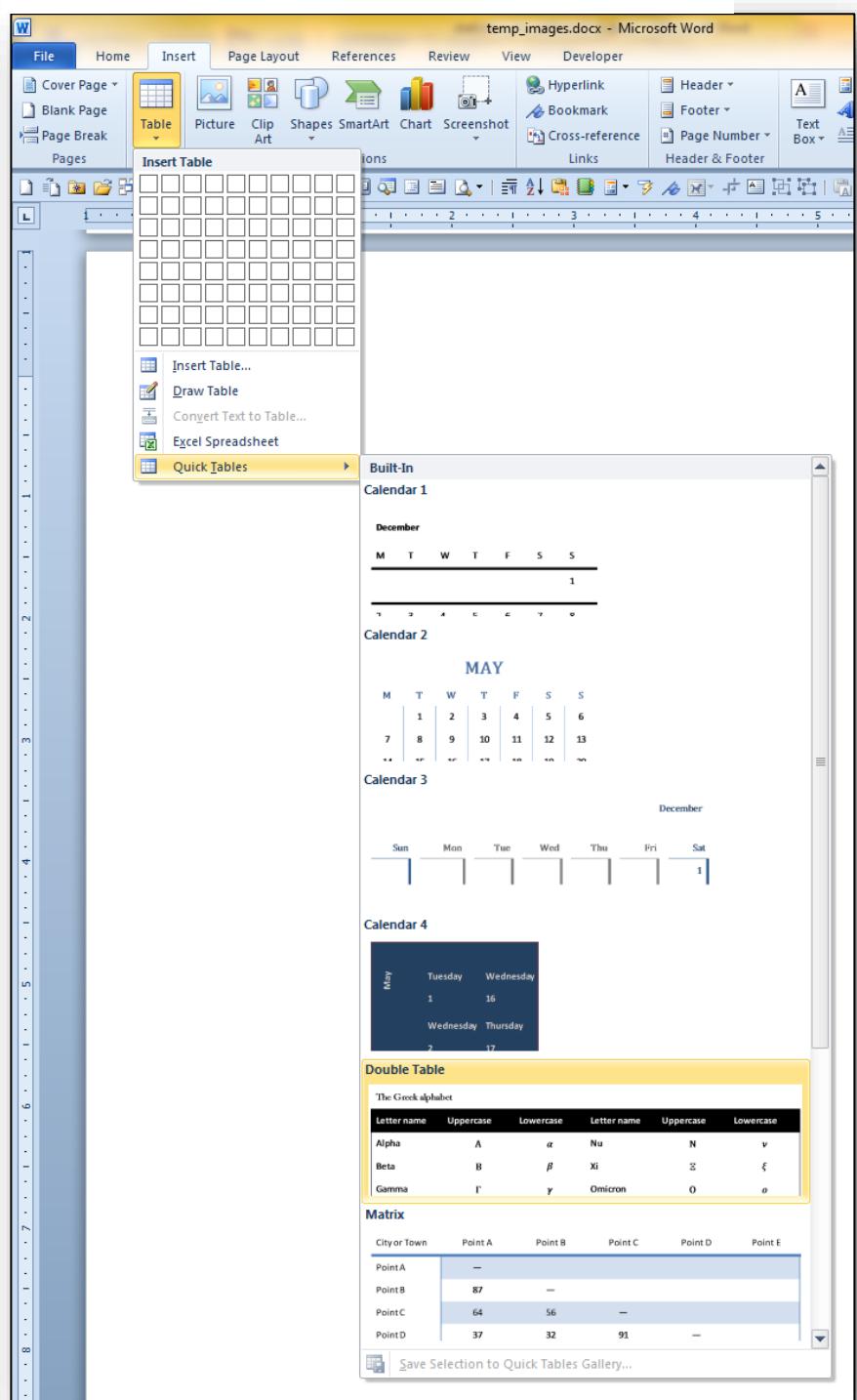
Other options shown in the menu in Figure 3-15 include the **Insert Table** function, which allows one to specify the number of rows and columns by entering those values into a form (shown in Figure 3-14).

Other functions in WORD include creating a table (**Convert Text to Table**) from a series of entries separated by the TAB character (the Tab key or **Alt** when inserting the character using the search-and-replace function)⁵¹ and creating an EXCEL spreadsheet directly in the WORD document without having to load EXCEL manually. Finally, the **Quick Tables** options offer a wide range of pre-formatted tables including calendars, double tables (splitting a table into two halves that appear side by side) and matrices (same categories on the top row and the left column).

⁵¹ The quickest way to access the search-and-replace menu is to use the key combination CTL-H.

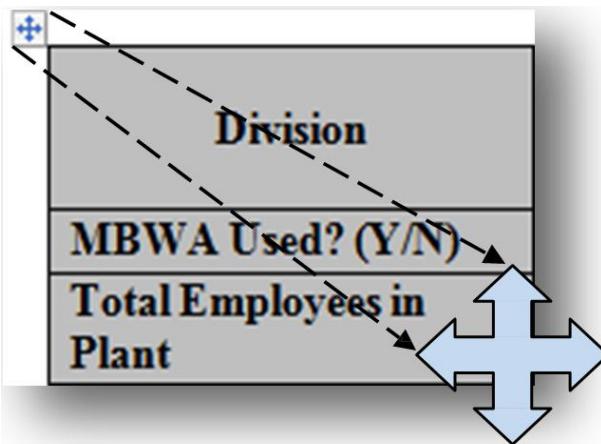
One can also save one's own customized table formats as “quick tables” in this gallery (last entry on the menu) by highlighting the desired table one has created and giving it a name.

Figure 3-16. Quick Tables in Excel.



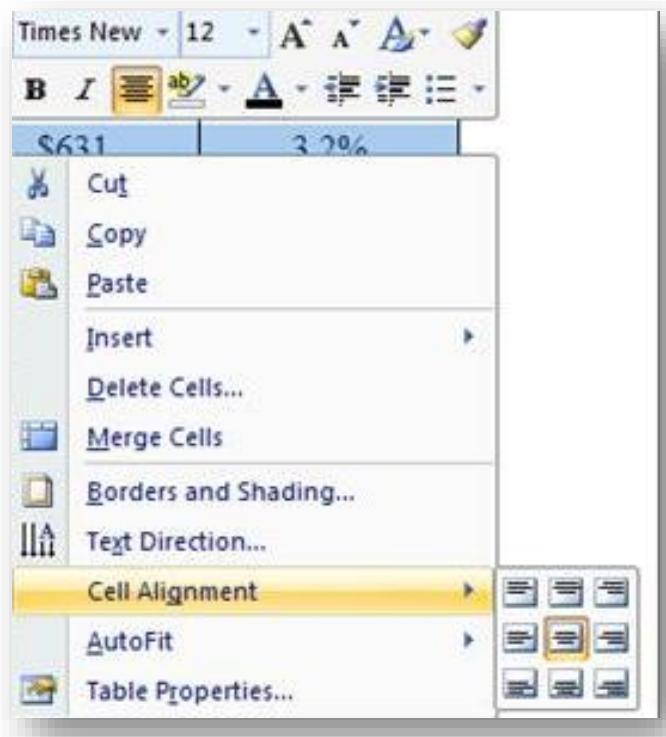
A particularly important set of tools in WORD is activated by right-clicking on an entire table using the four-arrow symbol in the left upper corner of the table as shown below in Figure 3-17.

Figure 3-17. Highlighting an entire table in Word.



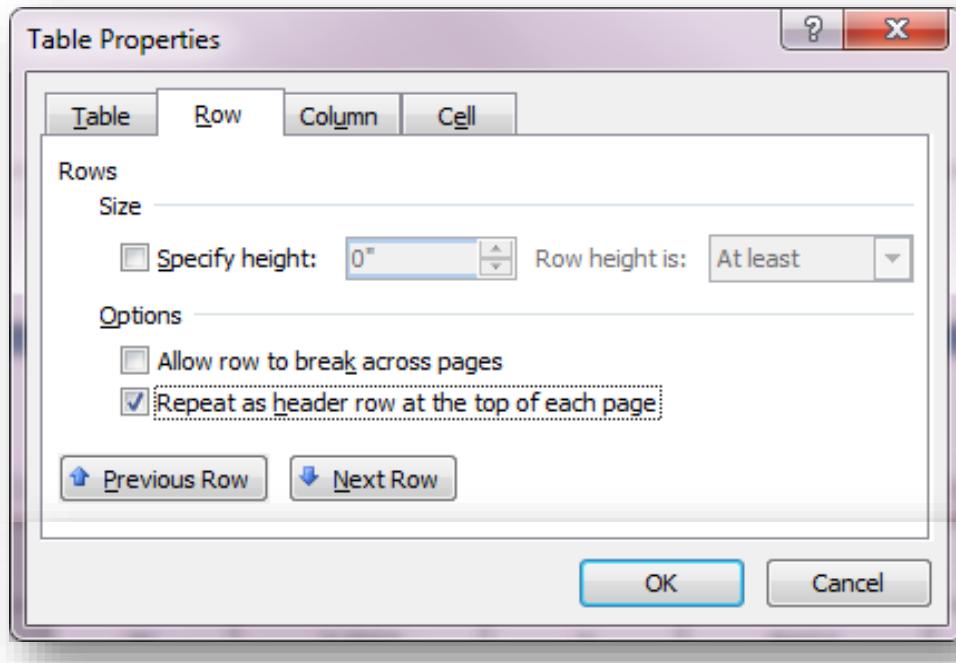
Highlighting the table, one or more rows or columns or a specific cell allows one to right-click for a formatting menu that includes **Borders** and **Shading** and **Cell Alignment** (illustrated in Figure 3-18). One can apply colors, shift the position of text left or right and up or down, and add colored borders to cells, rows, columns, or the entire table.

Figure 3-18. Right-click pop-up menu for adjusting formatting in Word table.



A useful option in **Table Properties** (the last option in Figure 3-18) for large tables is in the row option for the top rows that force the selected row(s) to be repeated at the top of every new page (see Figure 3-19) if the table is broken automatically into parts to fit the space available.

Figure 3-19. Table Properties menu showing header-row definition.



INSTANT TEST P 3-14

Practice creating a reasonably extensive table in Word - you may be able to use data from one of the Web sites you visited or material from a previous exercise. Even just half-a-dozen rows with three or four columns will do for the demonstration. Include a heading row on your table that labels the columns.

Move the table toward the bottom of a page in a Word document by inserting several Enter strokes before it and observe the data flowing automatically onto the next page. Unless you have already set the Table Properties to define the header, the data will not be labeled on the second page.

Now delete some empty lines before the table to bring it back and highlight the top row. Use the Table Options discussed in this section to force that top row to be used as a header on the next page if the table splits across a page-break.

Try the same process of pushing the table over the bottom of your page and note the difference in appearance of the second page when there's a heading row on every page.

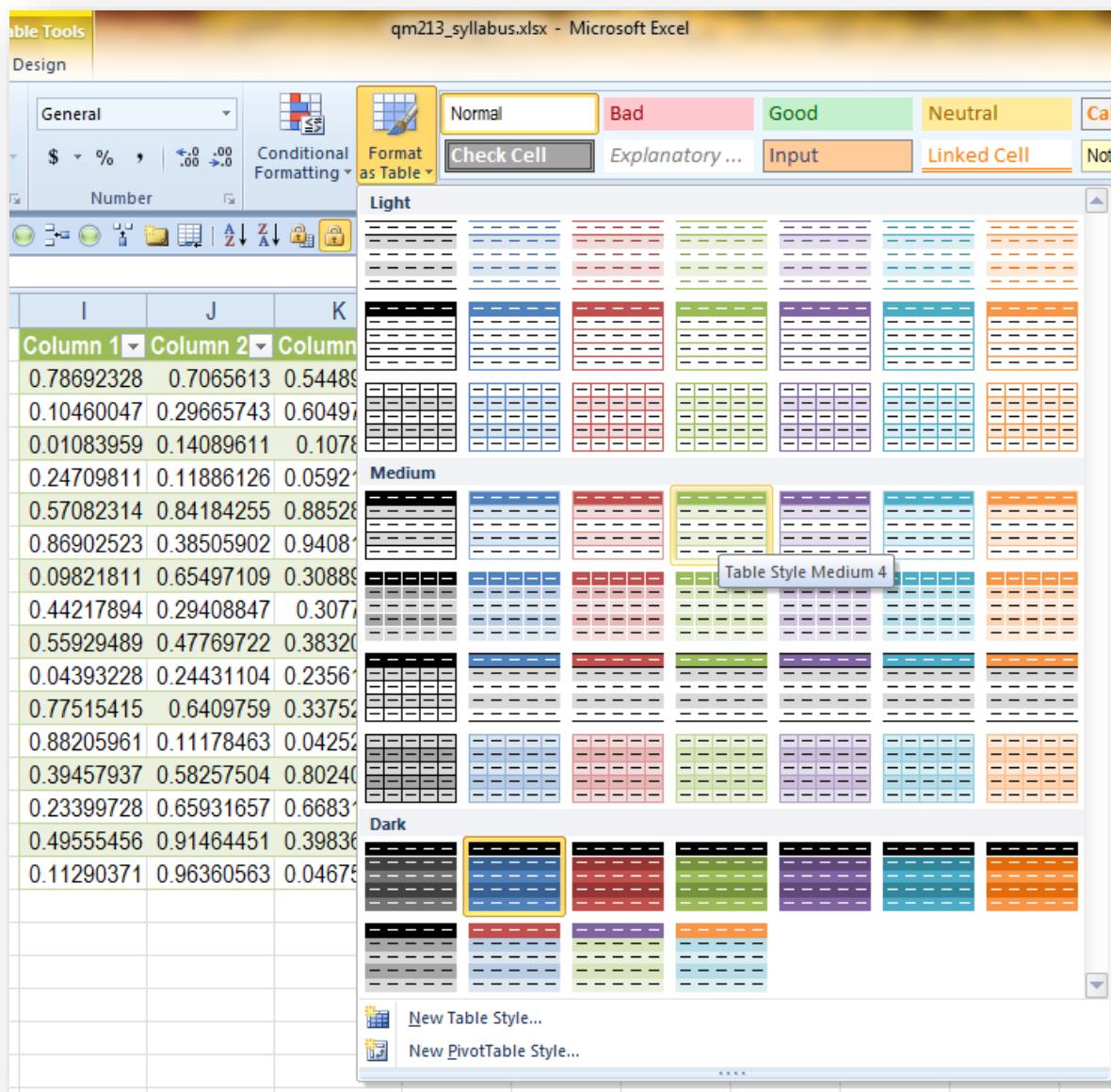
Be suitably impressed by your new-found abilities and give yourself a pat on the back.

If you *really* want to pick up hot dates at a bar,* learn to control whether a row transfers as a whole or slips across the page break line by line. Hint: use paragraph formatting. [*joke]

3.9 EXCEL Table Tools

EXCEL offers a much wider range of predefined table formats than WORD. One can do all the formatting oneself or use the **Format as Table** menu to choose among some popular color and emphasis layouts. A handy feature is that hovering the cursor over a style shows one the appearance of the selected table (see Figure 3-20). Notice that the selected format (row 4, column 4 in the choices shown) was applied temporarily to the table in columns I, J and K to illustrate its effect.

Figure 3-20. Trying out different table styles using Format as Table options.



INSTANT TEST P 3-15

Create a simple table in Excel and experiment by hovering your cursor over a variety of styles to see their effects. Click on one and then change it to another, and another, and another to get used the process.

3.10 Copying EXCEL Tables into a WORD Document

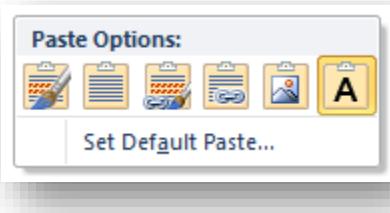
When copying an EXCEL table into a WORD document, there are several options. Figure 3-21 shows a sample table in EXCEL.

Figure 3-21. Sample Excel table for pasting into Word.

	A	B	C	D
1	Trans-Stellar Force	Sophonts	Starships	Weapons-class
2	Beth'dxomth'o	1,266,082	45,430	B
3	Cyphlonia	1,414,427	71,472	A
4	Mars	1,047,074	19,385	C
5	Raturto	981,772	63,954	A
6	Terra	860,295	4,977	B
7	Vakooli	377,009	61,488	D
8				

Simply Copying/Pasting (e.g., with CTL-C CTL-V or by using the **Copy** and **Paste** functions in the right-click menus) presents a menu for choosing how to paste the table (Figure 3-22) and also previews the appearance:

Figure 3-22. Paste menu.



The leftmost option, **Keep Source Formatting (K)**, pastes the EXCEL data as a table using the table formatting (e.g., column widths, emphasis such as bold) from the EXCEL source file into the WORD document. The K is the shortcut keystroke to select that option (instead of clicking on the icon). Figure 3-23 shows the preview:

Figure 3-23. Using the "Keep Source Formatting" option to paste Excel table into Word.

	Trans-Stellar Force	Sophonts	Starships	Weapons-class
	Beth'dxomth'o	1,266,082	45,430	B
	Cyphlonia	1,414,427	71,472	A
	Mars	1,047,074	19,385	C
	Raturto	981,772		
	Terra	860,295		
	Vakooli	377,009	61,488	D

Using the second icon from the left, **Use Destination Styles (S)**, puts the contents of the table into the document as a simple WORD table using whatever your defaults are for WORD tables (Figure 3-24):

Figure 3-24. Using the Destination Styles option to paste an EXCEL table into WORD.

Trans-Stellar Force	Sophonts	Starships	Weapons-class
Beth'dxomth'o	1,266,082	45,430	B
Cyphlonia	1,414,427	71,472	A
Mars	1,047,074	19,385	C
Raturno	981,772		
Terra	860,295		
Vakooli	377,009	61,488	D

The middle icons, **Link & Keep Source Formatting (F)** (Figure 3-25) and **Link & Use Destination Styles (L)** (Figure 3-26) bind your original EXCEL file to your WORD document.

Figure 3-25. Pasting an EXCEL table into WORD as a link using the source formatting.

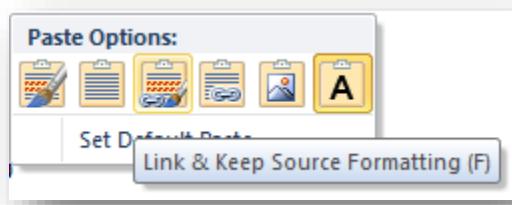
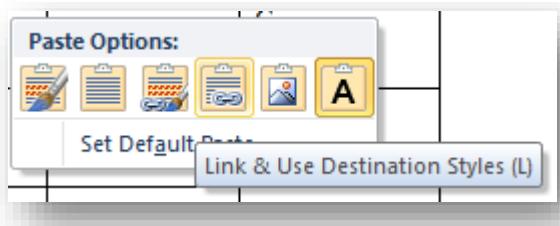


Figure 3-26. Pasting an EXCEL table into WORD as a link using the destination styles.



All numerical changes in the EXCEL source file are instantly reflected in the formatted view in your WORD document. The implication: you must keep track of both your EXCEL and your WORD files so that they never go out of synchronization (*synch*). If the EXCEL source file is moved from its original place into another folder or renamed, the table in WORD will disappear and any cross reference to it will show a highlighted message reading “**Error! Reference source not found.**” in the text.⁵²

Linked tables are ideal for periodic reports that show the current data every time you open them. Links to graphs (discussed later) do the same: whenever the source graph in the EXCEL file changes, the WORD report shows the modified graph the next time the WORD file is opened. All linkages also mean that mistakes can be propagated automatically from the EXCEL spreadsheet into all documents using the linkages.

For example, here (Figure 3-27) is the image of a dynamically linked table (using destination styles) in which the user has accidentally modified the last cell in the “Starships” column to the erroneous value “3” (red dot added to the image to identify the cell):

Figure 3-27. Dynamically linked table showing error propagating from Excel to Word.

Trans-Stellar Force	Sophonts	Starships	Weapons-class
Beth'dxomth'o	1,266,082	45,430	B
Cyphlonia	1,414,427	71,472	A
Mars	1,047,074	49,385	C
Raturto	981,772	63,954	A
Terra	860,295	4,977	B
Vakooli	377,009	3	D

INSTANT TEST P 3-18

Create a simple numerical table in Excel. Paste it into a Word document using the Link & Use Source Formatting option. Then do the same on another page of the same document using the Link & Use Destination Styles option. Compare the effects.

Now go back to your Excel source file and alter some of the numbers - and highlight them as colored text.

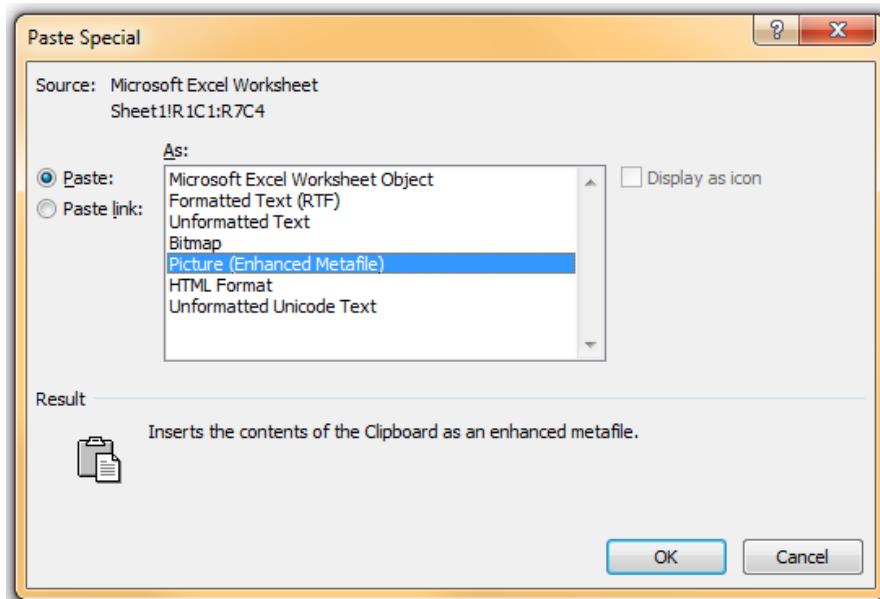
Go back to your Word document and see the effects on what's visible there.

⁵² To be sure everyone is clear on this, the “Error!...” string in the indicated line on this page does *not* indicate an error in *this* page of the text: it’s the text of the error message for illustration only. If you find such a warning elsewhere, let the instructor know.

The second option from the right in the Paste Options menu is Picture (U). This option (Figure 3-28) pastes an image of the EXCEL table into your document using your default preference for the type of image to use.

If you want more options for pasting images of an EXCEL table, use Ctrl-Alt-V to bring up more choices, as shown in the Paste Special menu in Figure 3-29:

Figure 3-29. Paste Special options.



Picture (Enhanced Metafile) and Bitmap both insert an exact image of what was in the EXCEL file, including the cell boundaries if they are visible in EXCEL. Bitmaps are much larger than enhanced metafiles, as you can see in the file listing in Figure 3-30:

Figure 3-28. Pasting an Excel table as a picture into Word.

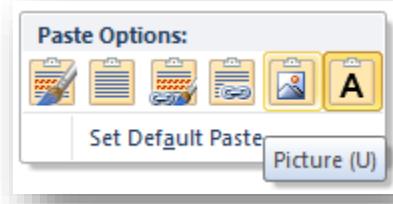


Figure 3-30. Comparing bitmap and enhanced metafile file sizes.

blank.docx	18 KB
blank_with_bitmap.docx	267 KB
blank_with_enhanced_metafile.docx	24 KB

INSTANT TEST P 3-19

Experiment with pasting different versions of an Excel table using the Paste Special menu. Try pasting the tables into a document with text you've already written to see the consequences for that text. **Remember to save your file as TEMP before you do any experimenting!**

4 Charts, Histograms, Errors in Graphing

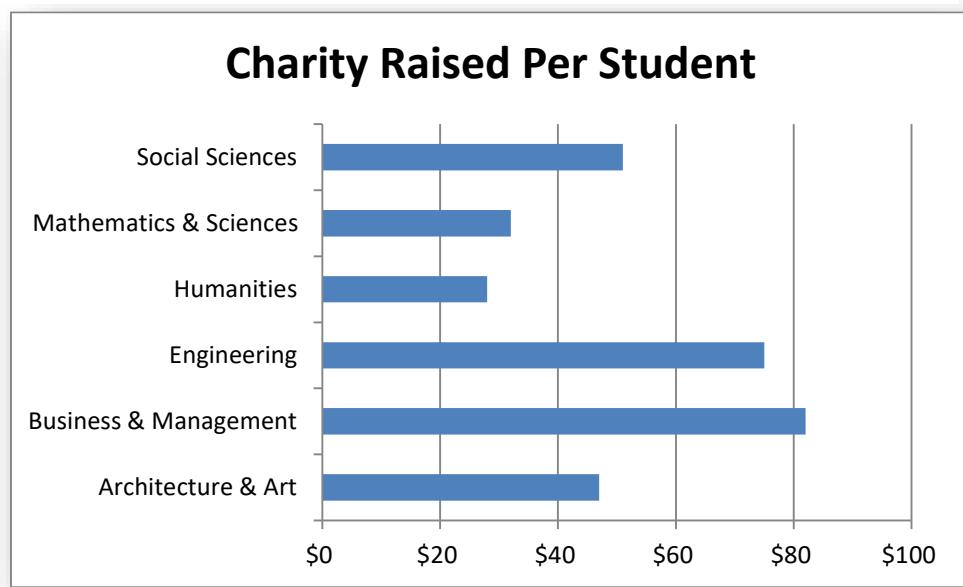
4.1 Horizontal Bar Charts vs Vertical Column Charts

When we need to think about and represent data sets involving categorical and ordinal data, we should use horizontal bar charts rather than vertical bar charts. Vertical bar charts are, by convention, usually reserved for interval and ratio scales. For example, consider the made-up data in about the average amount of charity donations raised by students in different schools of a university in Figure 4-1. A horizontal bar chart (Figure 4-2) represents the charity raised per student by the length of the horizontal bar corresponding to each school.

Figure 4-1. Donations per student by school.

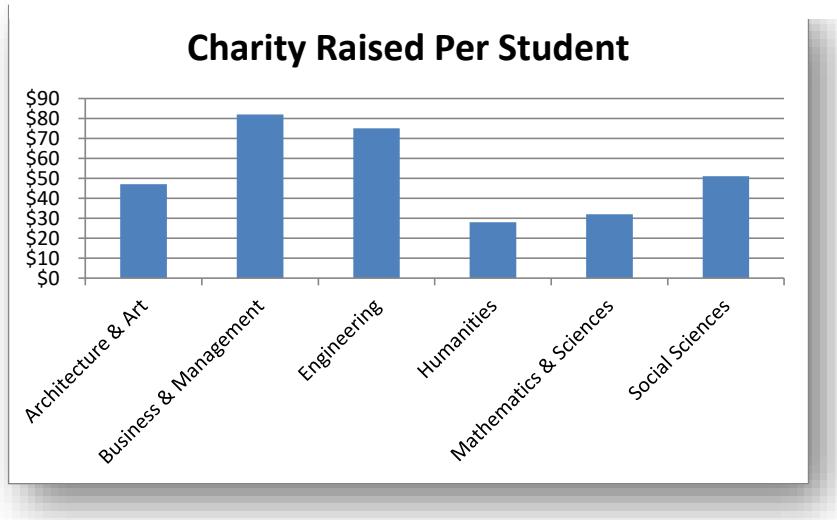
School	Charity Raised Per Student
Architecture & Art	\$47
Business & Management	\$82
Engineering	\$75
Humanities	\$28
Mathematics & Sciences	\$32
Social Sciences	\$51

Figure 4-2. Horizontal bar chart showing donations by school.



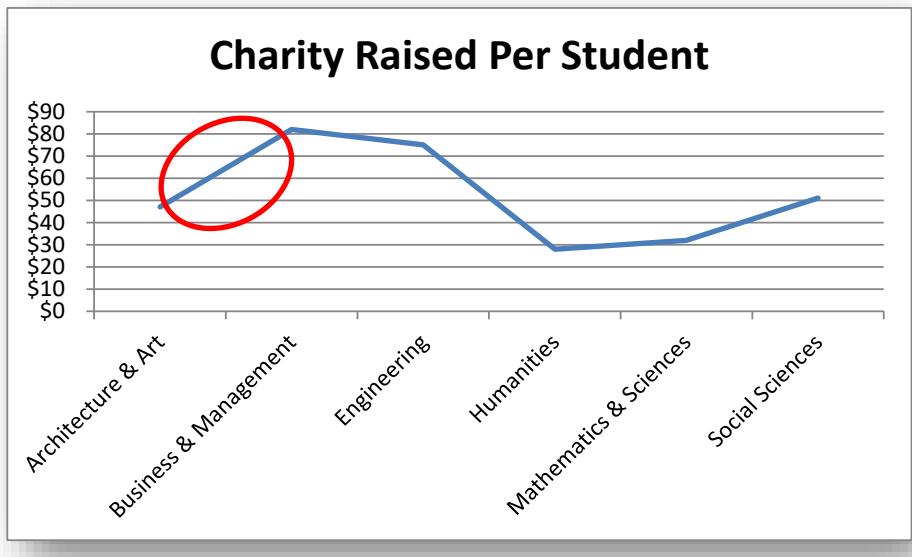
The advantage of horizontal bar charts versus vertical charts is that there is *no implication* that the axis with the category labels (in this case, the vertical axis) represents any kind of quantitative variable. In contrast, putting the bars vertically is (by a widely flouted convention) an indication that there is *some sort of quantitative relationship* among the labels of the bars. So although the same information *can* be represented as a vertical bar chart (Figure 4-3), some experts argue that the hidden assumption behind any charts that show the main variable(s) on the vertical axis (the *ordinate* or *Y-axis*) is that the horizontal axis (the *abscissa* or *X-axis*) is a continuous variable, providing some meaning to values between the observed markers.

Figure 4-3. Vertical bar (column) chart showing donations by school.



To illustrate how misleading it would be to put categorical variables on the X-axis, Figure 4-4 shows the same data as Figure 4-2 and Figure 4-3 but using a *line graph*. Everyone will agree that the values between the observed points (see red oval) make no sense at all: there are no values of the X-axis (donations per student) between values of the X-axis (school; e.g., *Architecture & Art* and *Business and Management*). Figure 4-4 is a definite no-no.

Figure 4-4. Bad choice of representation for categories.



4.2 Pie Charts

Another way of representing information about categories is the pie chart. Generally, a pie chart shows segments whose areas add up to a meaningful total and where the area of a segment (or what's equivalent, the angle of the vertex of the pie) corresponds proportionally to the value for that segment. Figure 4-5 shows the total charitable donations per school in a single semester at a university.

Figure 4-7 shows the pie graph for these data arranged with the segments in alphabetical order. Each segment

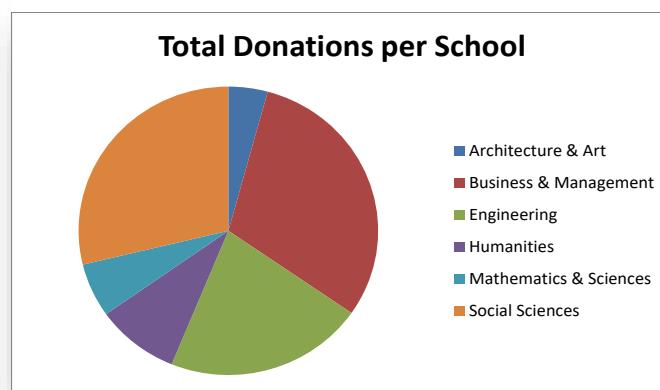
[Figure 4-5. Total donations per school in one semester.](#)

School	Total Donations
Architecture & Art	\$4,606
Business & Management	\$33,046
Engineering	\$23,401
Humanities	\$9,792
Mathematics & Sciences	\$6,560
Social Sciences	\$31,263

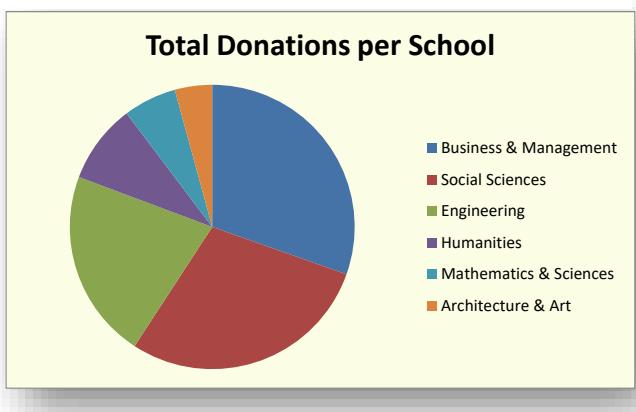
has a different color and is identified in the legend (the list of schools with a small colored square before each school).

One of the conventions that *may* be used in a pie chart is that the data are sorted by size and the largest component segment is placed starting at the 12:00 position, with the decreasing segments placed in clockwise order. Figure 4-6 shows the same data arranged in this conventional pattern.

[Figure 4-7. Pie chart showing total donations per school.](#)



[Figure 4-6. Pie-chart using clockwise pattern of descending values.](#)



To change chart type, right-click on the pie and use the **Change Chart Type** menu shown in Figure 4-8:

Figure 4-8. EXCEL menu for changing chart type.

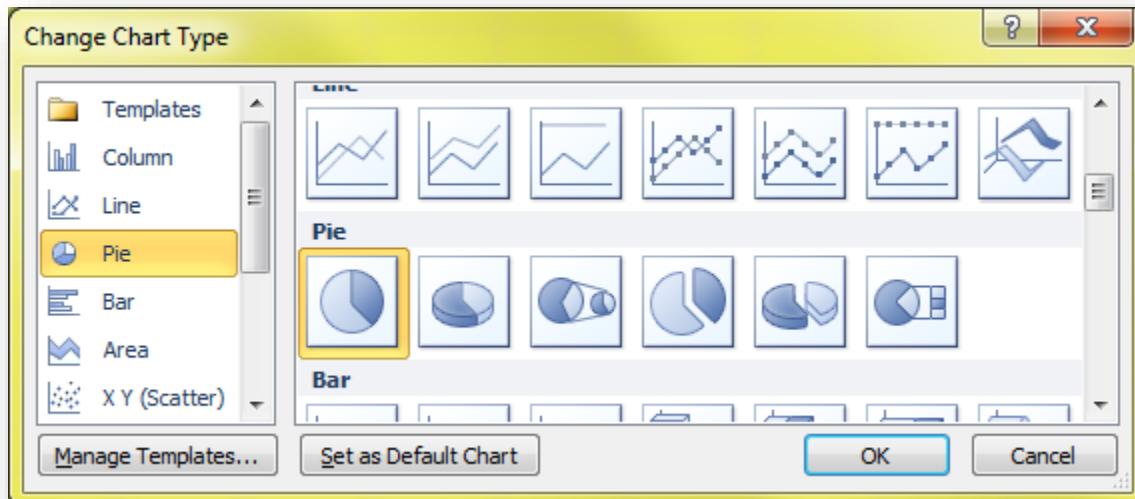
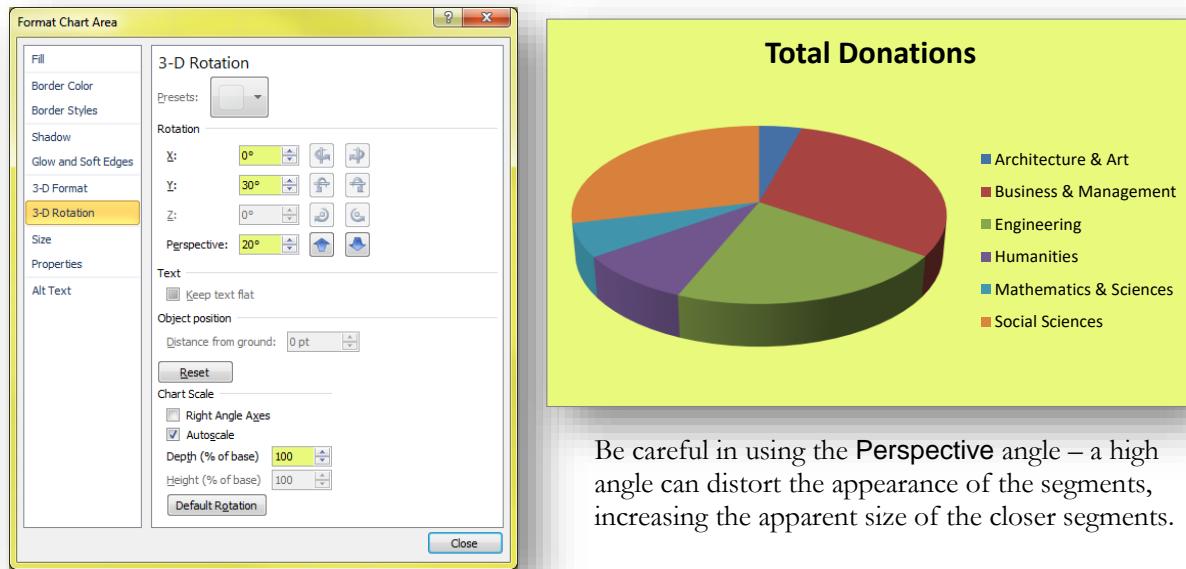


Figure 4-9 shows the result of converting Figure 4-7 to a 3D pie chart and then applying the 3-D Rotation options.

Figure 4-9. Menu and results for changing from pie chart to 3D pie chart and applying 3D rotation.

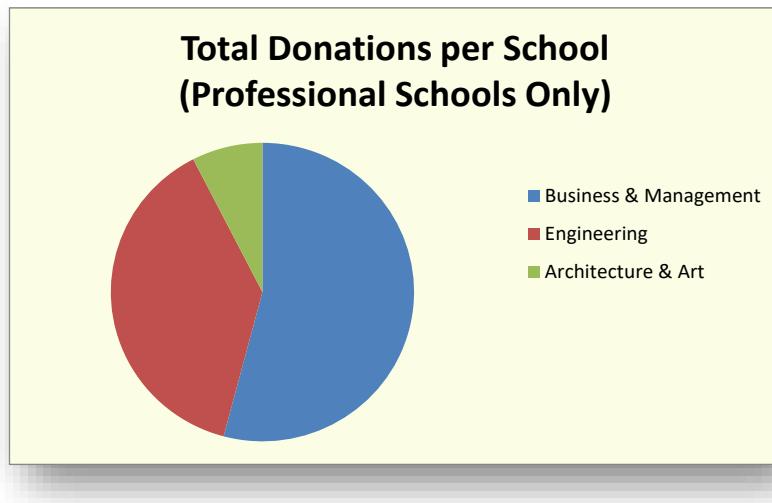


INSTANT TEST P 4

Create a pie chart with some interesting data. Change the style to 3D and experiment with various rotation angles in the X and Y axes using the 3-D Rotation menu. Try out options in the 3-D Format menu. Play around with other options such as Border Color, Border Styles, Shadow, and Glow and Soft Edges.

The total for a pie chart does not necessarily have to be the sum of *all* possible values in a table; it may be reasonable in specific cases to create a chart showing the relative proportions of *selected* components. For example, someone might be interested primarily in discussing the donations of students in the professional schools in the data of Figure 4-5; a pie chart for those selected data would look like Figure 4-10.

Figure 4-10. Total Donations per School (selected schools only).

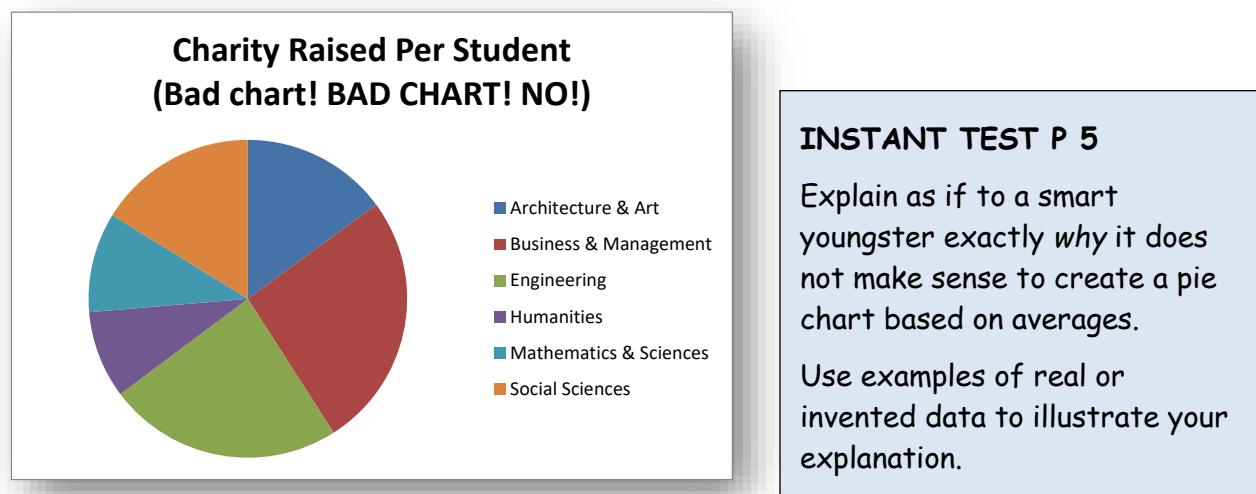


IMPORTANT WARNING ABOUT BAD PIE CHARTS

Don't create pie charts for data you cannot legitimately add up.

For example, Figure 4-11 is a nonsensical chart that shows donations *per student* for each of the schools. Those numbers are *averages*. You can't add up averages unless the number of data points (usually the *sample size*) are the same for all the averages. If the School of Architecture & Art has half the number of students that the School of Business & Management has, it doesn't make sense to add up their averages. So a pie chart of averages would not make sense, since one does not add averages up together – one computes the total from multiplying the average by the number of observations (students, here) that gave rise to the average. Figure 4-11 is an example of a *bad pie chart* that doesn't make sense.

Figure 4-11. Pie chart created using averages (BAD).



4.3 Clustered and Stacked Bar Charts and Column Charts

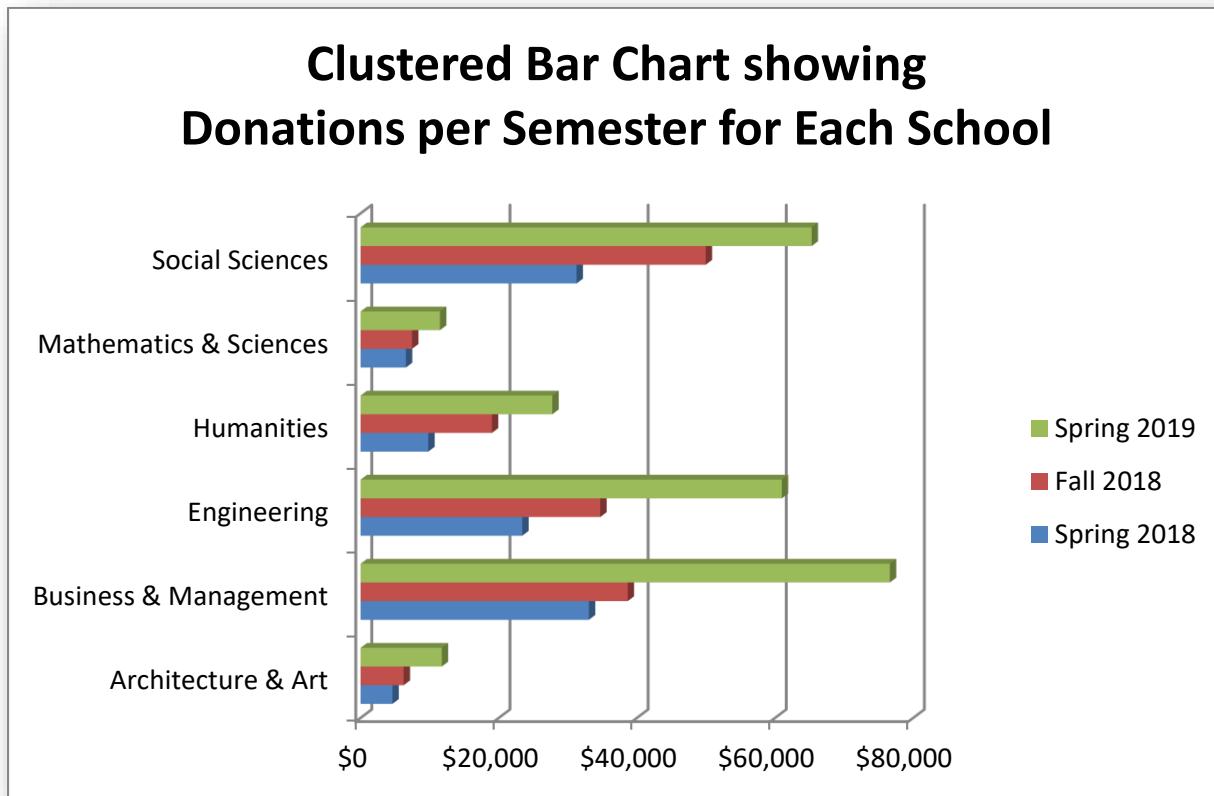
So far, we've been looking at charts for a single variable corresponding to each of several categories. However, we often have several variables for each value of the category; for example, Figure 4-12 shows the total charitable donations collected by each school in each of three semesters.

Figure 4-12. Total donations by school for each of three semesters.

School	Spring 2018	Fall 2018	Spring 2019
Architecture & Art	\$4,606	\$6,236	\$11,749
Business & Management	\$33,046	\$38,667	\$76,623
Engineering	\$23,401	\$34,695	\$60,979
Humanities	\$9,792	\$19,009	\$27,754
Mathematics & Sciences	\$6,560	\$7,444	\$11,488
Social Sciences	\$31,263	\$49,959	\$65,300

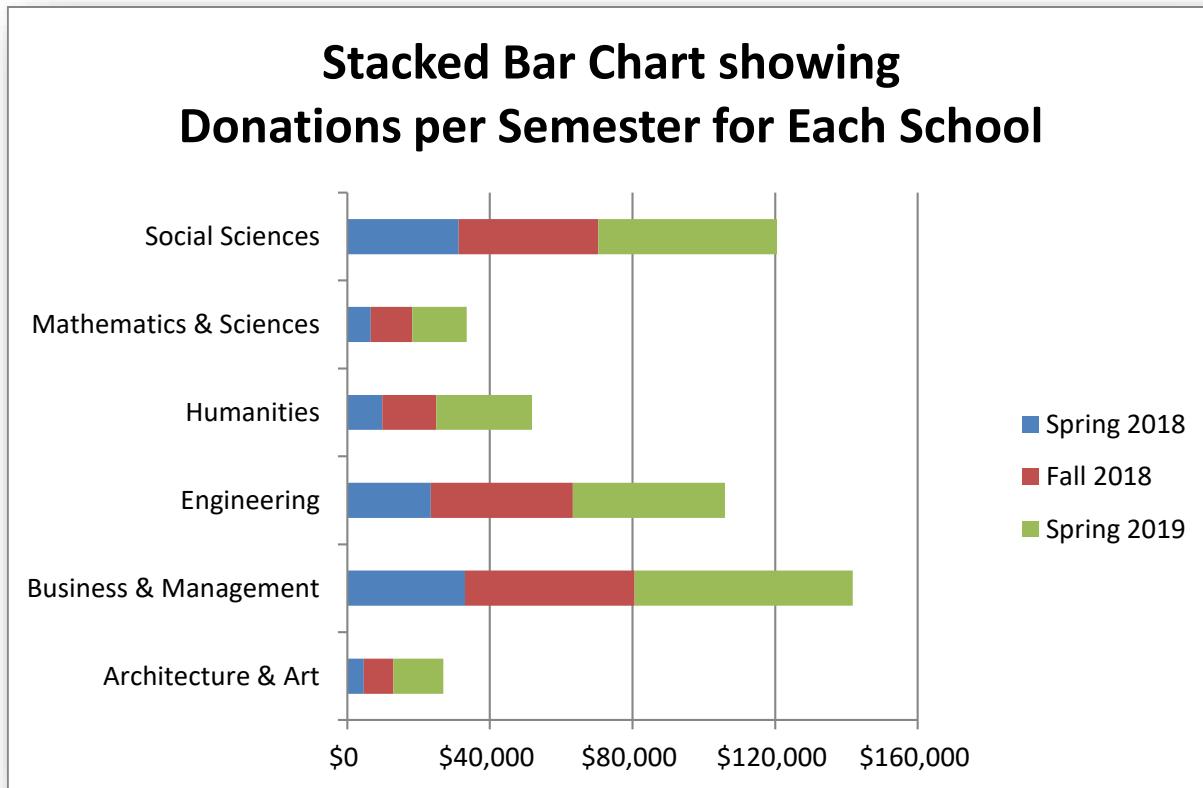
A *horizontal clustered bar chart* (Figure 4-13) can show each of the three semesters' totals as separate bars. It's easy to see the changes in donations for each of the schools over the three semesters in this display. It's also easy to compare the donations for each of the semesters for each of the six schools, as if the graph were a combination of three separate bar graphs overlaid to create a single picture.

Figure 4-13. Clustered horizontal bar chart.



An alternative for these data is to put the bars on top of each other in a *stacked bar chart*. Figure 4-14 shows the same data as in Figure 4-13, but it's much easier to evaluate the total donations for all three years lumped together for each school. In addition, it's also easier to see the relative size of the yearly donations within one school's record.

Figure 4-14. Stacked bar chart.

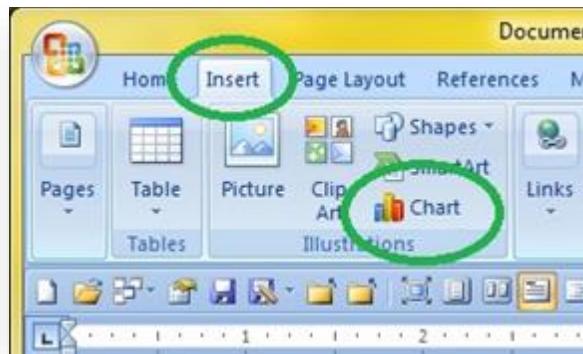


As you can see, the different layouts affect the perception and focus of the viewer. Take the time to decide which kind of format makes the most sense for your purposes. You may even want to represent the same data in various ways as your discussion moves from one issue to another when presenting your findings about a phenomenon. But an important insight is that there's no rigid rule that tells you "always use this" or "always use that;" you have to *think about your purpose* when you decide how to graph your data.

4.4 Creating Charts in WORD

It is possible to create a chart in WORD, but the menu functions (Figure 4-15) immediately transfer control to EXCEL using a dummy table that you have to modify to suit your needs (Figure 4-16).

Figure 4-15. Word *Insert Chart* menu functions.



It makes much more sense to start your own EXCEL process, enter your data, and use the EXCEL chart tools to create your graph. You can then copy and paste the graph in whatever form (Enhanced metafile, etc.) suits your needs. For simplicity, use the **Top and Bottom** option in the **Wrap Text** options and be aware that sometimes you have to apply that option to the Figure or Table caption as well.

Figure 4-16. Dummy Excel table instantiated by Word *Insert Chart* function.

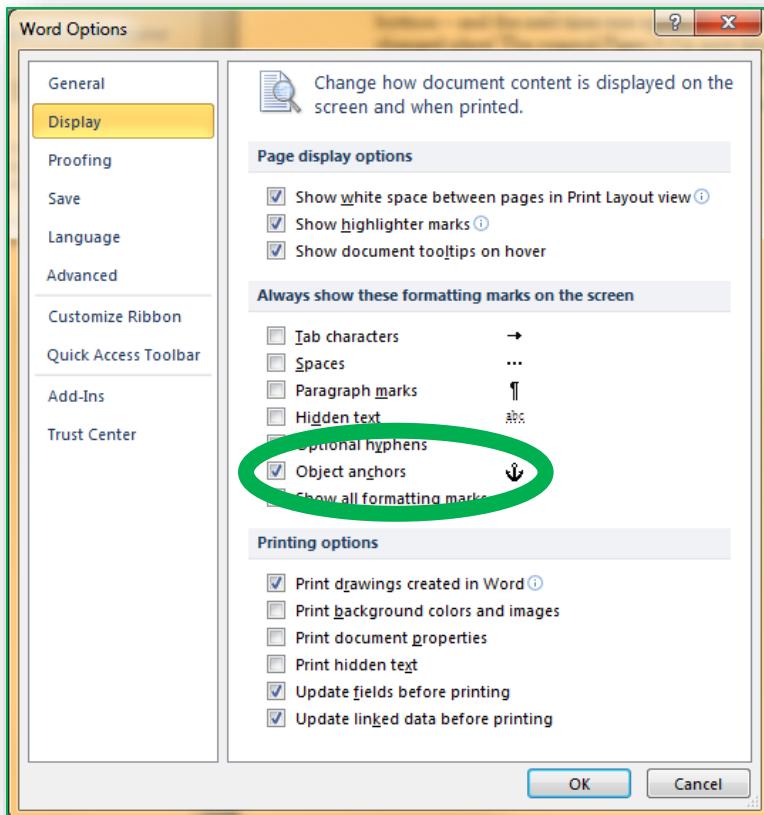
A screenshot of Microsoft Excel. The ribbon shows the 'File' tab is selected. The 'Home' tab is active. A table is selected, spanning from cell A6 to D10. The data in the table is as follows:

	A	B	C	D	E	F	G
1		Series 1	Series 2	Series 3			
2	Category 1	4.3	2.4	2			
3	Category 2	2.5	4.4	2			
4	Category 3	3.5	1.8	3			
5	Category 4	4.5	2.8	5			
6							
7							
8	To resize chart data range, drag lower right corner of range.						
9							

4.5 Managing Figure & Table Numbers in WORD

One of the most annoying features of pasting graphics into a WORD document is that their identifying numbers sometimes go out of order. One can have a *Figure 3-2* at the top of a page and *Figure 3-3* at the bottom – and the next time one opens the file, the figures are still in the same place but their labels have changed place! The original *Figure 3-2* is now labeled *Figure 3-3* and has the wrong description; the original *Figure 3-3* is now labeled *Figure 3-2* and has the wrong description. Cross-references become scrambled, too. The solution is to go to the **File | Options** menu and check the **Object anchors** box, as shown in Figure 4-17.

Figure 4-17. Options for showing object anchors.

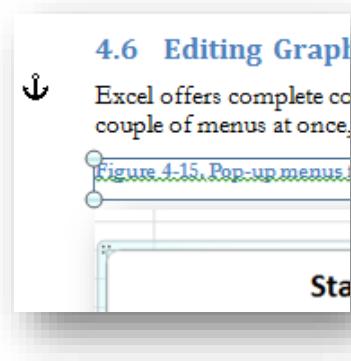


If labels, figures or tables are out of order, click on the top of the label or of the image to make WORD show you where the anchor is located (Figure 4-18), then drag the anchor to the appropriate position (paragraph) in the text ensure that the numbers are in the right order. This process also helps in positioning tables that seem eager to escape onto the next page or who jump on top of another figure.

If all else fails, you can cut an image from your document, paste it into a temporary blank document, remove the caption, and then start again. You have to remove inline cross-references when you do this and insert them again too to avoid the “Error!” message in your text.

Finally, a useful trick for updating all cross-references is to highlight the entire document (**Ctrl-A**) and then press **F9**. Using the **Print Preview** function also works.

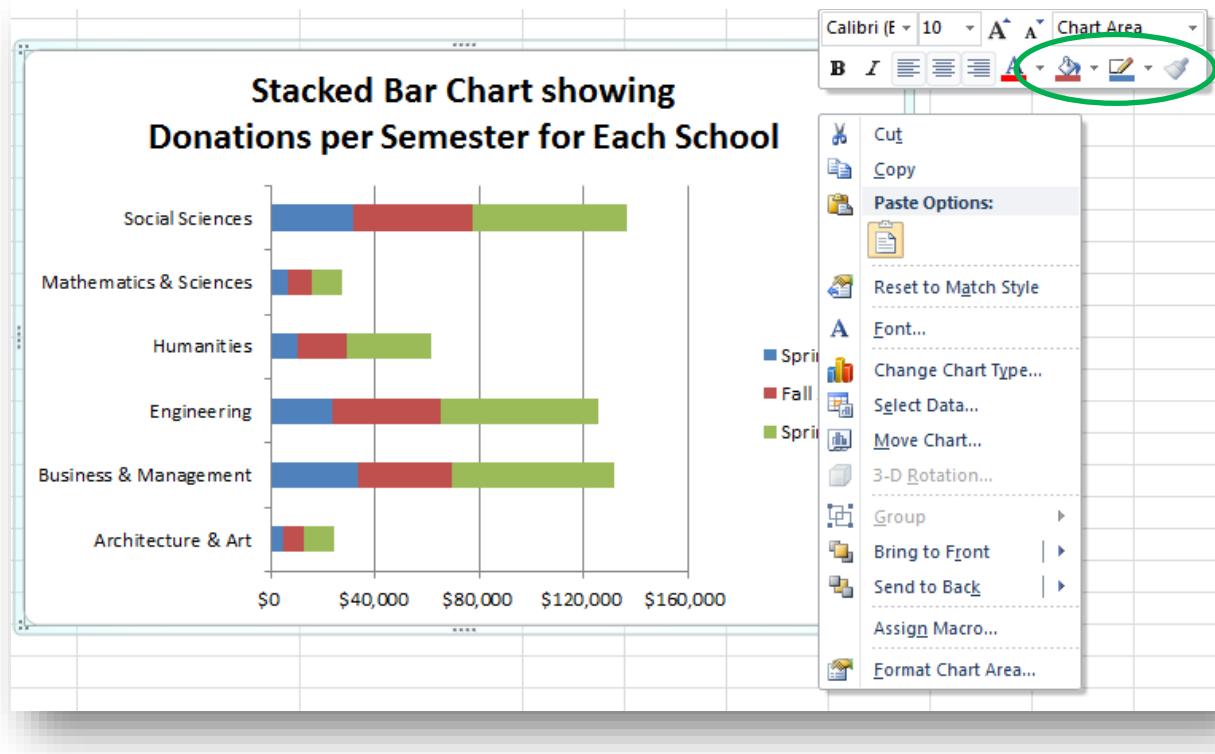
Figure 4-18. Anchor point visible.



4.6 Editing Graphics in EXCEL

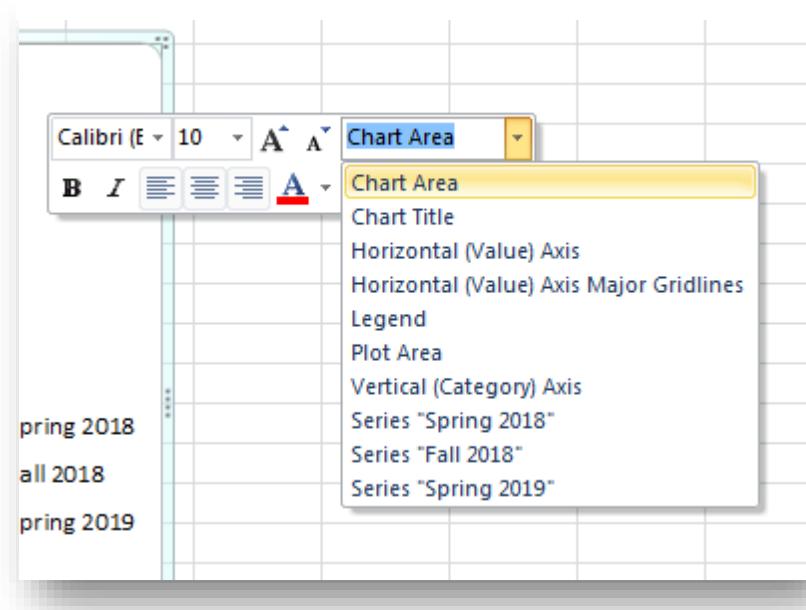
EXCEL offers complete control over every element of graphs. Right-clicking anywhere on a chart brings up a couple of menus at once, as shown in Figure 4-19.

Figure 4-19. Pop-up menus for editing charts in Excel 2010.



The pull-down menu expands as shown in Figure 4-20.

Figure 4-20. Pull-down menu for editing charts in Excel 2010.



4.7 Frequency Distributions

We often count the occurrences or frequencies of specific values in our observations. For example, we might want to summarize information about customer satisfaction shown in the following data set in Figure 4-21 (only the top and bottom of the data table are shown in this figure).

The frequency distribution in Figure 4-22 lists the number of observations that are equal to or *smaller* than the value in the row label (e.g., 0, 10, 20...) and greater than the value of the next-lower row label (except in the first row). Thus there are no observations of customer satisfaction less than 40; there are 6 entries with values between 41 and 50 inclusive, 26 entries with values between 51 and 60 inclusive, and so on. There are 5 customer-satisfaction values between 91 and 100.

An important guideline when creating histograms is that categories with fewer than 5 entries can distort statistical calculations. To be clear, **there's nothing wrong with reporting the exact counts** in the categories, but under some circumstances you may need to combine adjacent categories to reach the minimum frequency of 5. **If there are such categories, you can combine adjacent categories until you reach a minimum of 5.** For example, in Figure 4-22, if there had hypothetically been 3 observations in the 21-30 category and 2 in the 31-40 category, we could have defined a 21-40 category labeled 40 with 5 observations in all.

Figure 4-21. Data set with 200 customer-satisfaction data.

Customer Satisfaction (0-100)	
59	
78	
69	
81	
72	
75	
71	
81	
66	
81	
70	
80	
75	

Figure 4-22. Frequency distribution of customer-satisfaction data.

Customer Satisfaction (0-100)	Frequency
0	0
10	0
20	0
30	0
40	0
50	6
60	26
70	74
80	61
90	28
100	5

A bin boundary indicates the number of observations from just above the previous bin to the value of the bin.

Thus there were 26 values between 51 and 60 inclusive in the data represented in this frequency distribution.

4.8 Histograms

The graphical representation of the frequency distribution is called a *histogram* and is generally a vertical bar chart, as shown in Figure 4-23.

Figure 4-23. Histogram.



INSTANT TEST P 12

Find some interesting real frequency data and graph them using pie charts, clustered bar charts, and stacked bar charts. Go one step further and use a 100% Stacked Bar and decide what it tells you. Come up with a set of guidelines on what determines your choice among these options (What are you trying to illustrate?) and post your thoughts in the Discussion area in NUoodle for this week.

4.9 Creating Frequency Distributions and Histograms in EXCEL

Using the Data | Data Analysis |Histogram function of EXCEL, one can convert a set of data into a frequency distribution, organize the classes by rank, and create a simple chart – all automatically. Figure 4-24 shows the initial menu items for activation of the selection of statistical tools.

Figure 4-24. Initial sequence to access statistical tools in Excel 2010.

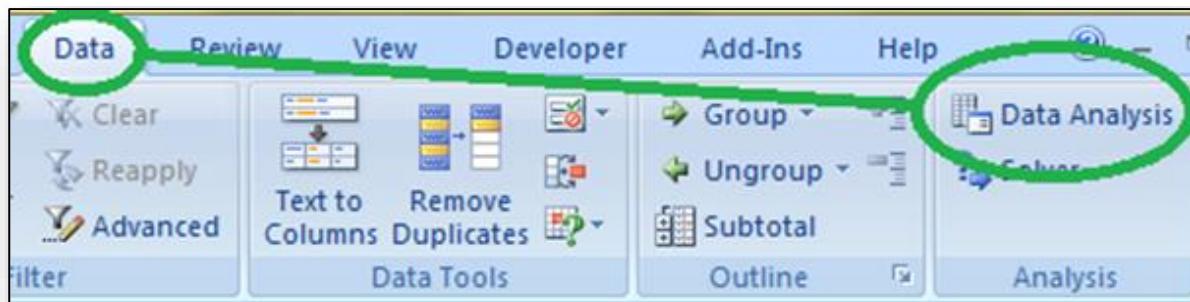


Figure 4-25. Data Analysis pop-up menu in Excel showing Histogram tool.

Clicking on the **Data Analysis** button shown in Figure 4-24 brings up the selection of **Analysis Tools** shown in Figure 4-25.

There are a total of 19 tools available; Figure 4-26 shows the remaining Analysis Tools.

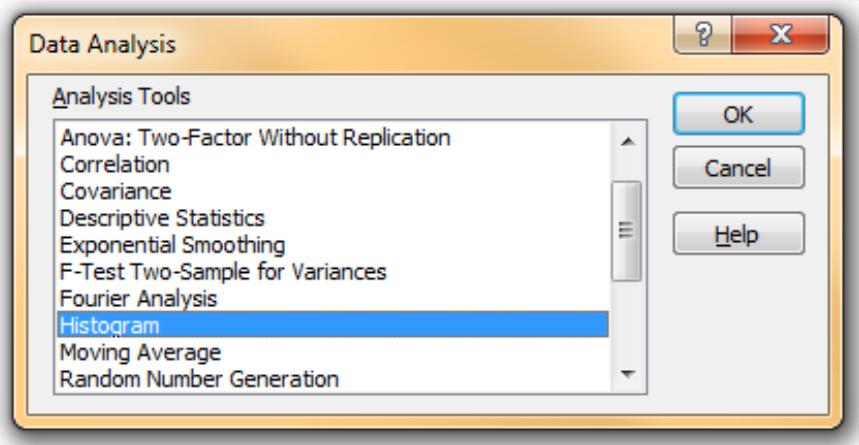
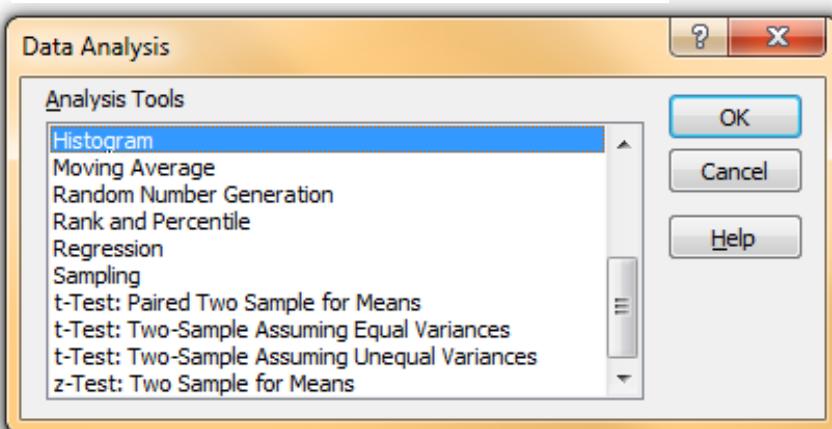
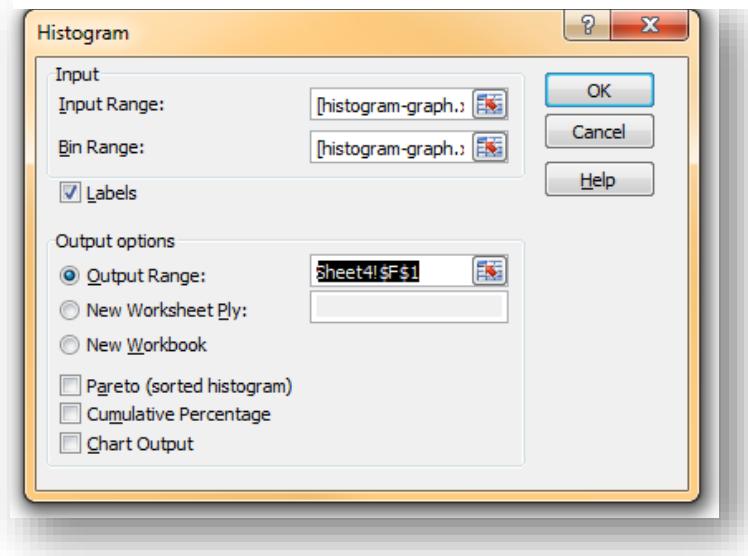


Figure 4-26. Remaining Analysis Tools.



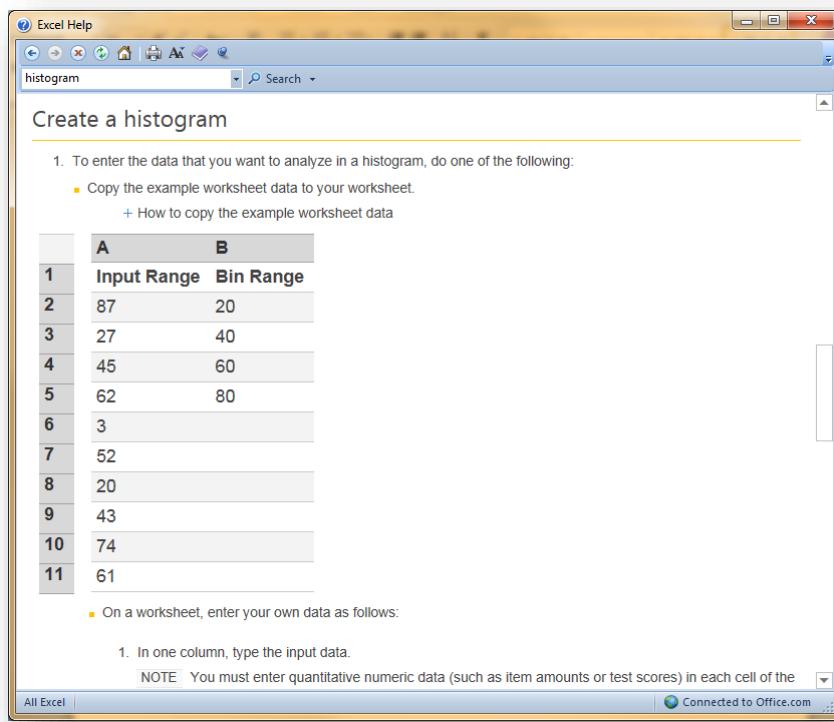
The pop-up menu shown in Figure 4-27 allows one to point to the columns – you can have multiple columns in the same chart – for input. The **Bin Range** stipulates where you have defined the categories you want EXCEL to tally. The **Output Range** (or **New Worksheet Ply** or **New Workbook**) offer options on where to put the tallies.

Figure 4-27. A *Histogram* tool pop-up menu.



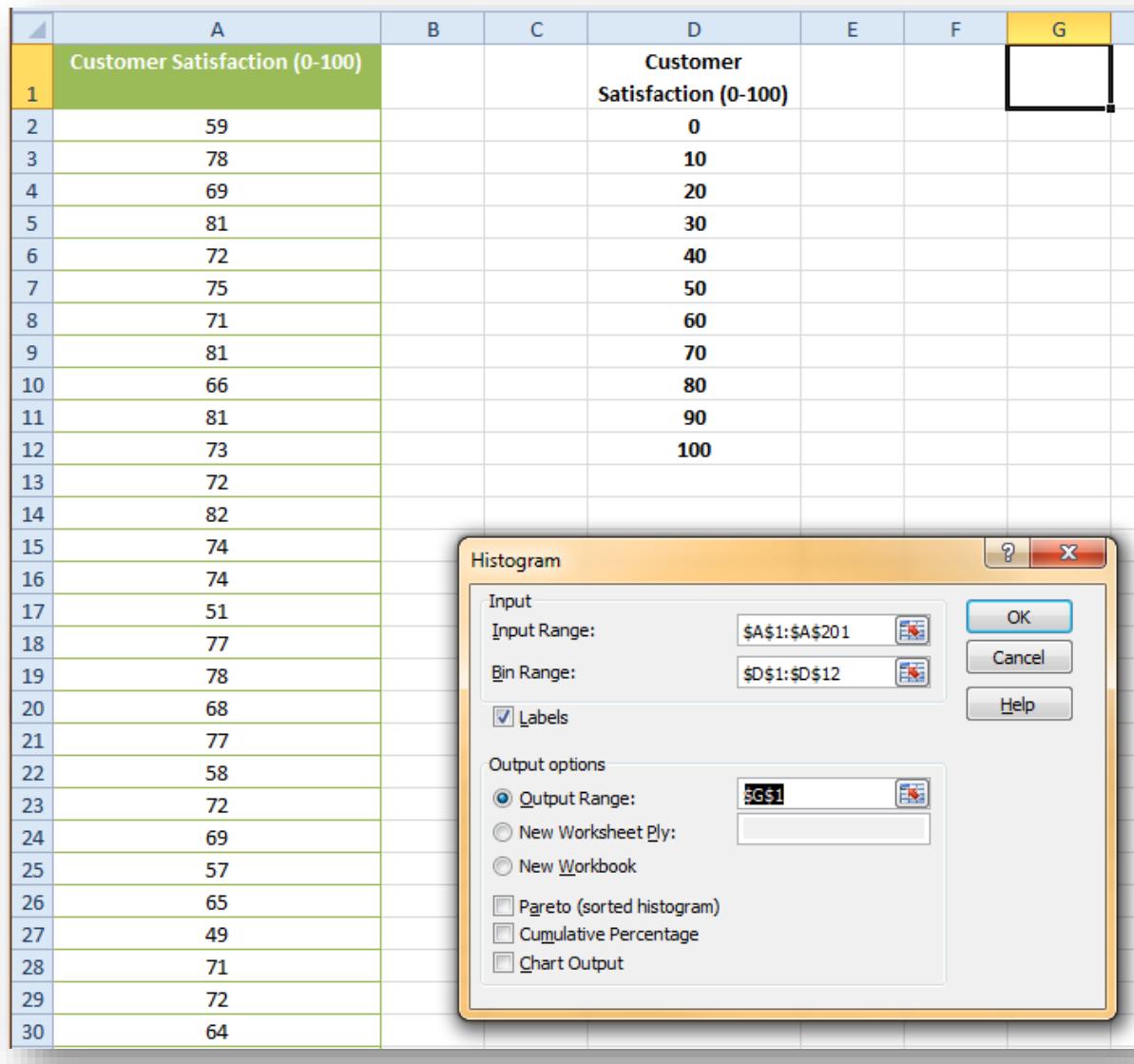
All the details of the options are available in the Help function (Figure 4-28).

Figure 4-28. Help for Histograms tool.



Be careful when enter the ranges in the menu shown in Figure 4-29. When the user clicks on the **Output Range** button, a glitch in programming switches the cursor back to **Input Range**. Highlighting the desired Output Range then wipes out the Input Range, to the consternation of the user and sometimes much bad language.

Figure 4-29. Defining the Input Range, Bin Range, use of Labels and Output Range for Histogram in pop-up menu.



INSTANT TEST P 4-15

Practice using the Histogram function. See what happens when you click on Output Range - observe where your cursor goes. Use the tool on some data you have acquired and see what happens as you increase the number of columns. Find out what error you get if you include the top row of labels but fail to check the Labels box.

Clicking on OK generates the frequency distribution (EXCEL calls it the **histogram**) in the desired location as shown in Figure 4-31. The counts show the number of cases greater than the lower Bin and *up to and including* the Bin value, as shown in Figure 4-31.

Figure 4-30. Table generated by *Histogram* Analysis Tool.

G	H
Customer Satisfaction (0-100)	Frequency
0	0
10	0
20	0
30	0
40	0
50	6
60	26
70	74
80	61
90	28
100	5
More	0

Figure 4-31. Demonstrating definition of bins.

A		B	C
1	DATA	Bin	Frequency
2	2.0	0	0
3	2.1	1	0
4	4.0	2	1
5	4.0	3	1
6	4.5	4	2
7	5.0	5	2
8	6.0	6	1
9	6.2	7	2
10	7.0	8	1
11	8.0	9	1
12	9.0	10	1
13	10.0	More	0

The options available for the **Histogram** function (Figure 4-29) are as follows:

- The **Input Range** and **Bin Range** are where one indicates the data areas. The data may be arranged in a single column or in a single row.
- **Bin Range** refers to a list of categories defined by the user; in this example, 0 to 100 in divisions of 10 seemed appropriate. The **Bin Range** must be arranged in a single column.
- Checking the **Labels** box allows the ranges to include descriptive labels.
- The **Output Range** can point to the same worksheet, as in this example, or it can be created in a new worksheet in the same workbook file (.XLSX) – or even into a new workbook.
- The option for **Pareto (sorted histogram)** generates additional columns showing the bins sorted downward by frequency (*modal* – most frequent – bin at the top).
- **Cumulative Percentage** computes the cumulative relative frequency distribution (discussed below).
- **Chart Output** produces a simple vertical histogram showing the frequency distribution and its graph (called an ogive, also discussed below).

INSTANT TEST P 4-16

Experiment with the options at the bottom of the **Histogram** menu. See what happens when you click on the various choices. Put each new histogram in a new worksheet play.

Experiment with using the **New Workbook** option as well.

4.10 Choosing a Reasonable Number of X-axis Values

When creating tables and charts with a numerical abscissa, how many classes should you use?

Sometimes the number of classes for which we have collected data becomes unwieldy. Imagine that we are studying the effects of a marketing campaign on the percentage of returned items in 1,250 retail stores? How could we reasonably present a table or a graph showing each individual store's results in the study? Even if we used multiple columns, such a massive table could take up several pages and would result in mind-numbing detail for our readers. Similarly, a graph with 1,250 categories on the abscissa would take up several arm-lengths of space if the names of each store were to be included.

A solution is *grouping*. We could classify the stores according to some reasonable criterion⁵³ such as geographical location (if all the stores are in the USA, perhaps state would be a reasonable basis for grouping, or maybe areas such as “northeast, southeast, north central, south central, northwest, and southwest).

Alternatively, an analyst might want to focus on the size of the stores and group the results by the total gross revenues; e.g., \$1M to \$4.9M, \$5M to \$9.9M, and so on.⁵⁴

Some guidelines for grouping data for statistical representation and analysis:

- In general, *somewhere between 10 and 30 groups* seems reasonable for tabular and graphical presentations.
- If your grouping criterion is numerical (e.g., total revenue, number of employees, number of shares sold, level of production errors) you can estimate the optimal interval size by calculating the range (the largest value minus the smallest value) and then dividing the range by 10 and also by 30.
- For example, if you have share prices as the criterion for grouping companies in an analysis, and the cheapest share costs \$138 whereas the most expensive share costs \$3,848, then the range would be $\$3848 - \$138 = \$3710$.
- So the smallest interval (the one producing 30 groups) would be $\$3710/30 = 123.7$ or about 124.
- The largest interval (the one producing 10 groups) would be $\$3710/10 = \371 .
- Avoid peculiar groups such as 27.3:37.2; groups starting and ending on 0s and 5s are common (e.g., 100, 105, 110... or 1200, 1400, 1600...). You don't have to start at the minimum, either; if your minimum were 128 and your groups were 20 units wide, you could start with the first group at 120:139, the second at 140:160 and so on.
- You might pick something like an interval of \$200 or \$250 to keep things neat; that would generate $\$3710/\$200 = 18.5$ groups which means 19 groups in all (you can't have a fractional group). The \$250 interval would produce 15 groups.
- The first \$200-wide group in this example could thus start off being \$0 to 199; the second group by share price would be \$200:\$399; the third, \$400:\$599. The last group under this scheme would be \$3800:\$3999.
- Avoid categories with fewer than five observations. Combine adjacent categories if necessary to arrive at a minimum of 5 occurrences per category.
 - You may want to sort your initial groups by the number of data and then combine adjacent low-frequency groups until you reach the minimum of five observations per group.
 - For example, if there were only two entries in the \$0:\$199 group described above and six in the \$200:\$399 group, you could combine those two into a \$0:\$399 group and have eight in it.

⁵³ A *criterion* is a basis for judgement. The plural is *criteria*. Don't say or write “the criteria is” or “the criterion are.” From Greek κριτήριον, *kriterion* from κρίτης, *krites* = judge. Our words *critic* and *criticism* come from the same root.

⁵⁴ Groups are usually indicated using colons (\$0:\$199) or ellipses (\$0..\$199) but not dashes (\$0-\$199). We avoid hyphens and dashes to avoid ambiguity with subtraction when negative numbers are included in the ranges (e.g., -15 --10, which looks like a mistake).

4.11 Problems with Disparate Quantities

Suppose the Urgonian Corporation carried out a survey of customer satisfaction in seven markets around the solar system. Figure 4-32 shows the results in terms of raw data and of percentage of positive responses. The layout of the chart is *vertical bar* only to save space on the page: usually it would be horizontal bars.

If we try to represent the raw data about the *number* of positive responses in a histogram (Figure 4-33), we immediately run into trouble: the range of total responses is so large (from 180,554 down to 77) that it is impossible to show anything meaningful about the lower frequencies on the same

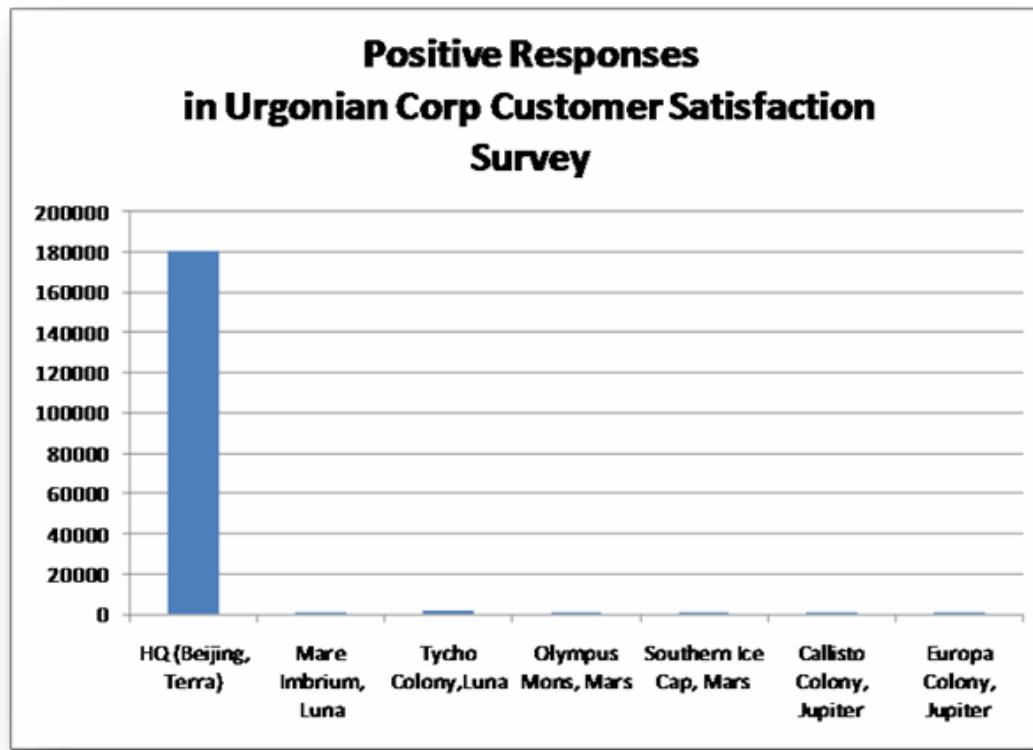
graph. All we can tell is that there were a lot of responses from Beijing; details of all the other sites are obscured because they are so tiny compared with the total from Beijing.

In general, it is not useful to try to show wildly different quantities on the same graph using a linear scale for the ordinate if the maximum is more than about 20 times the minimum. In the data for Figure 4-33, the largest value is *2,345 times larger* than the smallest value.

Figure 4-32. Urgonian Corporation survey results.

SURVEYS 2218.04.22	Returned	Positive	% Positive
HQ (Beijing, Terra)	324,932	180554	56%
Mare Imbrium, Luna	1847	1231	67%
Tycho Colony, Luna	2950	1915	65%
Olympus Mons, Mars	372	278	75%
Southern Ice Cap, Mars	414	266	64%
Calisto Colony, Jupiter	112	77	69%
Europa Colony, Jupiter	185	137	74%

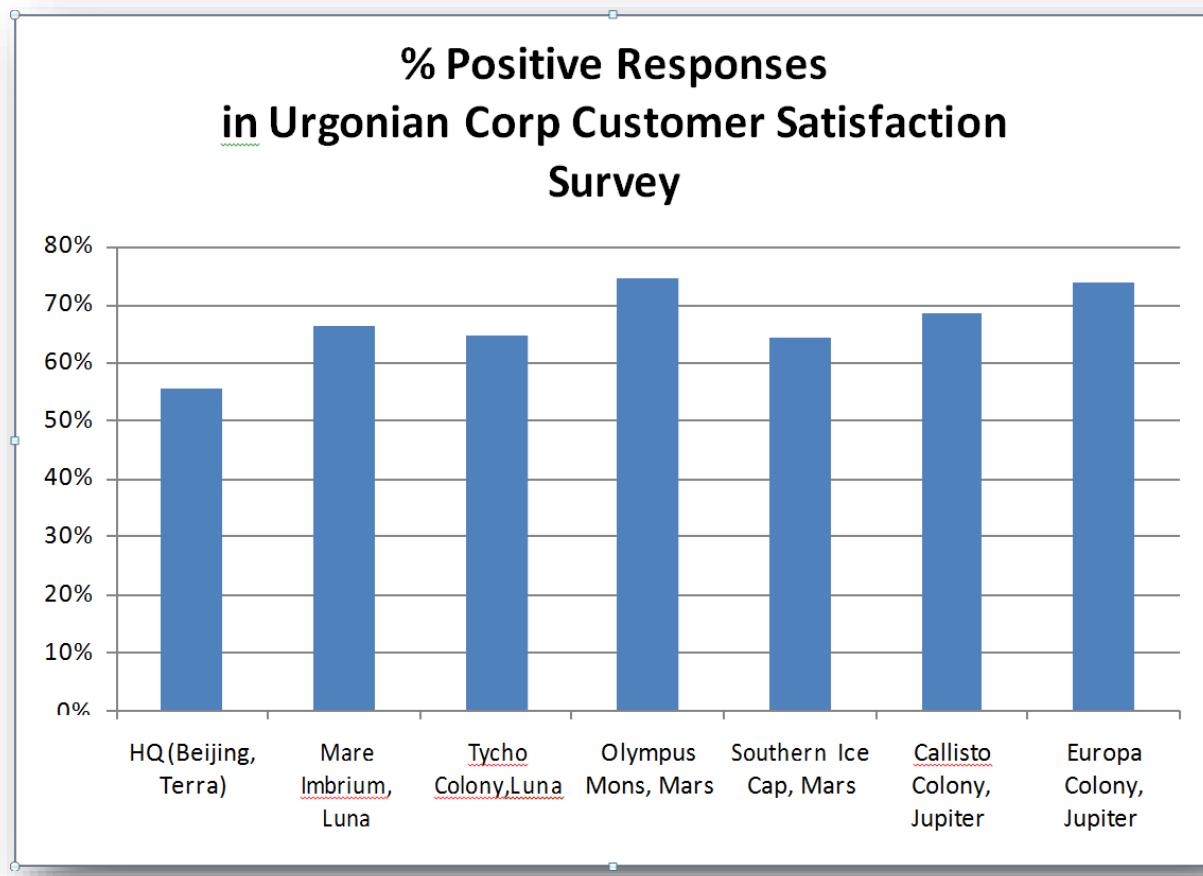
Figure 4-33. Useless vertical bar chart with disparate values.



Second, showing only the positive returns tells us nothing about the relative *proportion* of positive returns, which are more likely to be interesting. We must always think about whether a representation is meaningful before taking our time to create it.

Figure 4-34 shows the percentages of positive returns for the Urgonian Corporation customer satisfaction survey. Since those percentages vary from 56% to 75% and the theoretical limits are 0% and 100%, we should be OK.

Figure 4-34. Vertical bar chart of percentages of positive responses.



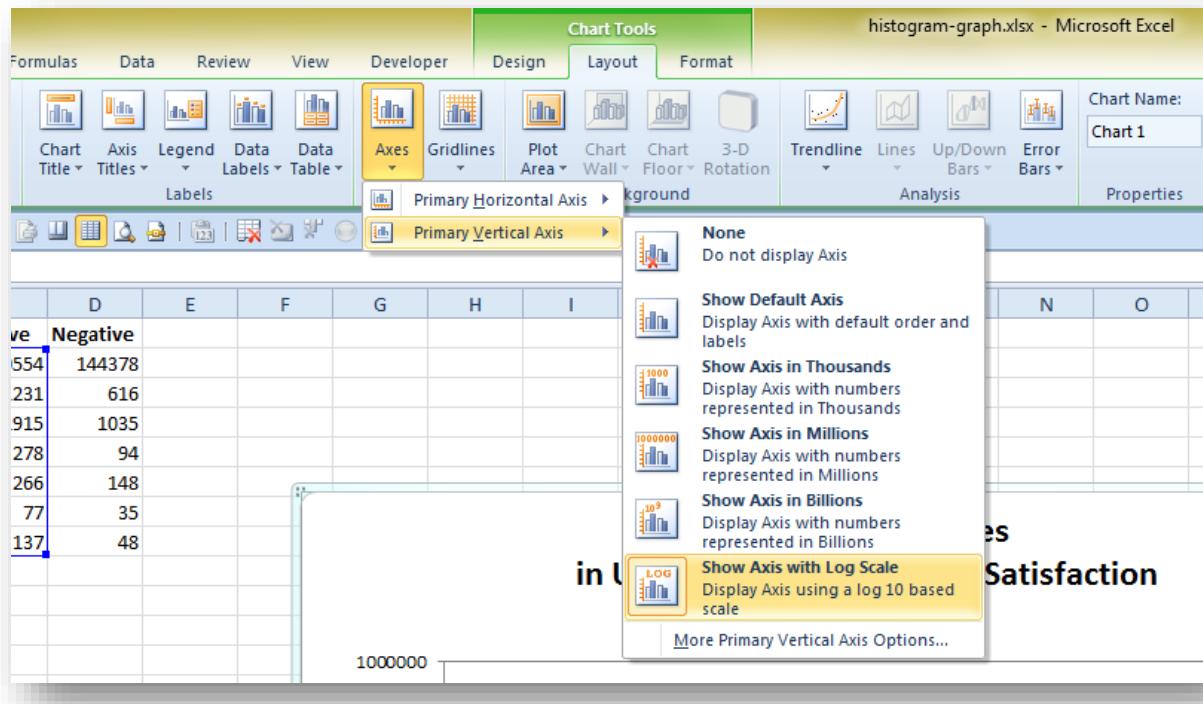
INSTANT TEST P 4-19

Experiment with the data in Figure 4-32 to create horizontal bar charts with these values. Try out different options such as the shape of the bars (3D, cones...) and examine the effects of options such as those for the Y-axis scale.

4.12 Logarithmic Scale on the Ordinate

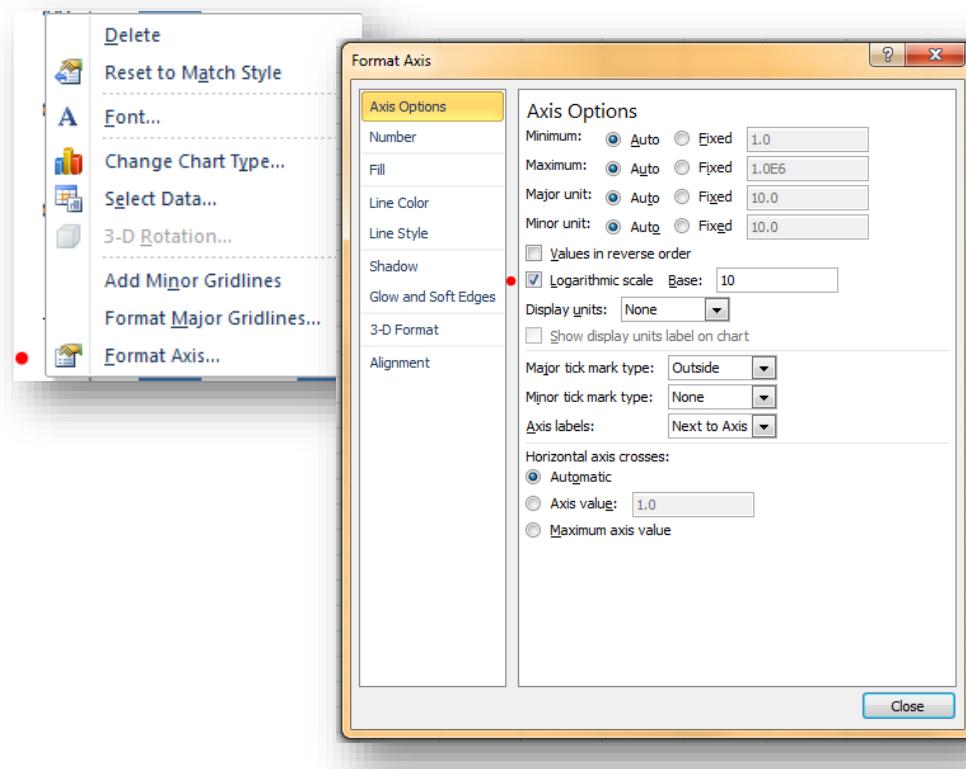
If you still need to represent the original disparate data on the Y-axis, you can use the logarithmic-scale option shown in Figure 4-35.

Figure 4-35. Excel *Chart Tools* menus to define logarithmic scale on ordinate.



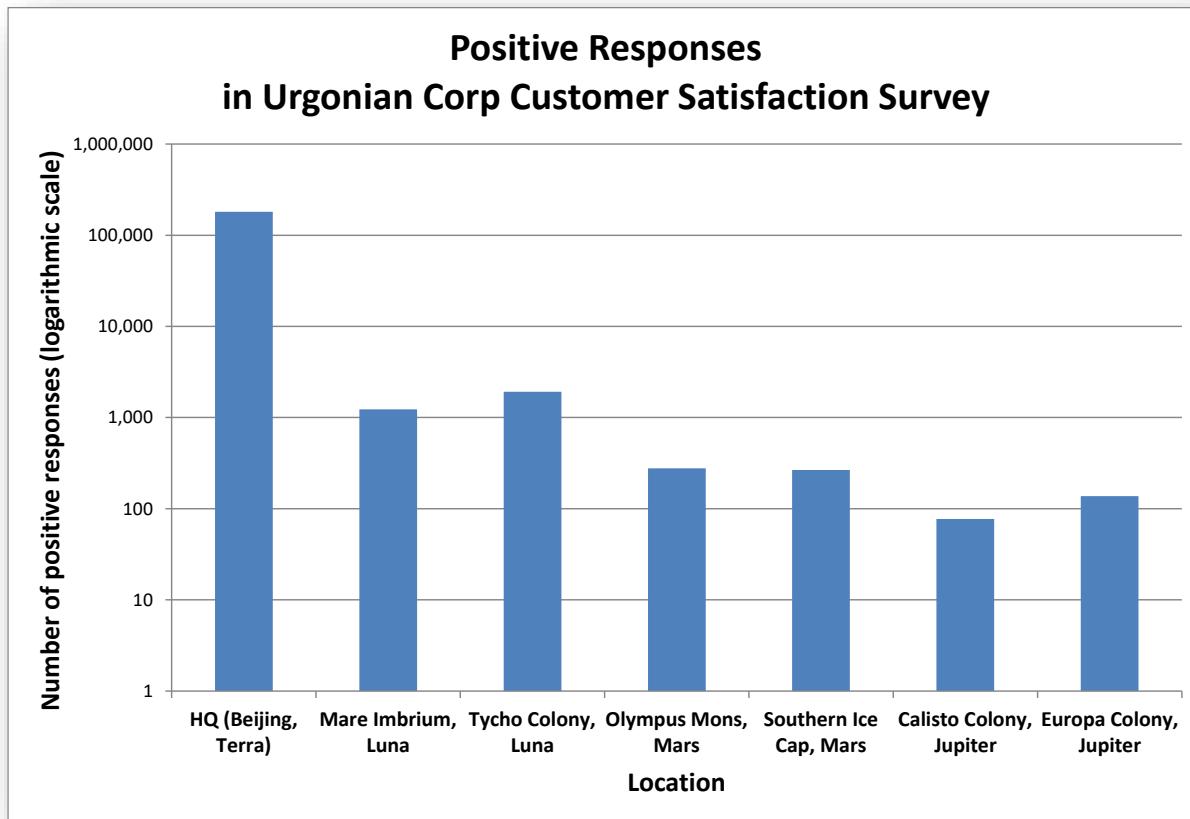
An alternative is to right-click on the ordinate scale to bring up the pop-up menu and select the **Format Axis...** function to bring up the **Format Axis** pop-up menu as shown in Figure 4-36.

Figure 4-36. Right-click menu for vertical axis and *Format Axis* pop-up.



The results are shown in Figure 4-37. Notice that the ordinate (Y-axis) is correctly identified as using a logarithmic scale.

Figure 4-37. Urgonian data with log scale Y-axis.



A word of warning: log scales make perfect sense to people who are familiar with them but can be confusing for people unfamiliar with the concept of logarithms.

- For example, a value that is twice as large as another on a \log_{10} scale (as in the example here) is ten times larger in reality.
- The vertical bar for the number of positive responses from the Mare Imbrium site is half the size of the bar for the data from HQ (Beijing) – but there are 100 times more positive responses in Beijing than in Mare Imbrium.

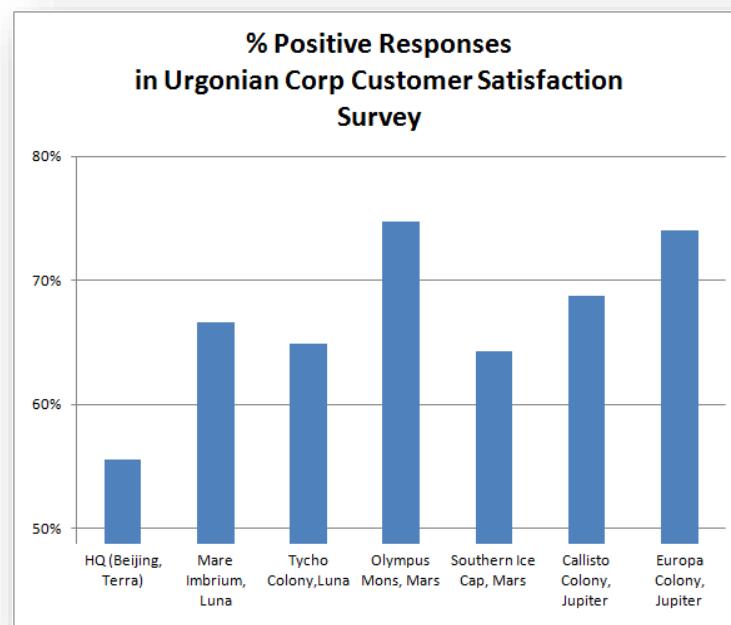
Some additional recommendations:

- If you use such a graph with naïve users – people unfamiliar with the very concept of a logarithm – there is a risk that they will misinterpret the data.
- On a nastier note, beware anyone who uses a log scale without explicitly mentioning it; they may be trying to trick the user of the graph into a misunderstanding.
- Finally, don't fall into the error of labeling the Y-axis as "Log of [whatever it is]." If we were to label the values with the actual logarithms in Figure 4-37, the Y-axis labels would read (from the bottom) 0, 1, 2, 3 and so on.

4.13 Truncating the Ordinate

One of the most serious errors that beginners (or dishonest professionals) can make is to inflate the apparent differences among groups in a histogram by cutting off the bottom. For example, the distortion in Figure 4-38 cuts off the bottom 50% of the histogram ordinate and grossly distorts the differences among the divisions of the corporation. A casual viewer might think that the Olympus Mons division has a positive rating several times that of the Beijing division, but in fact the difference is not nearly as large as the impression created by the dishonest representation.

Figure 4-38. Survey results with truncated Y-axis.



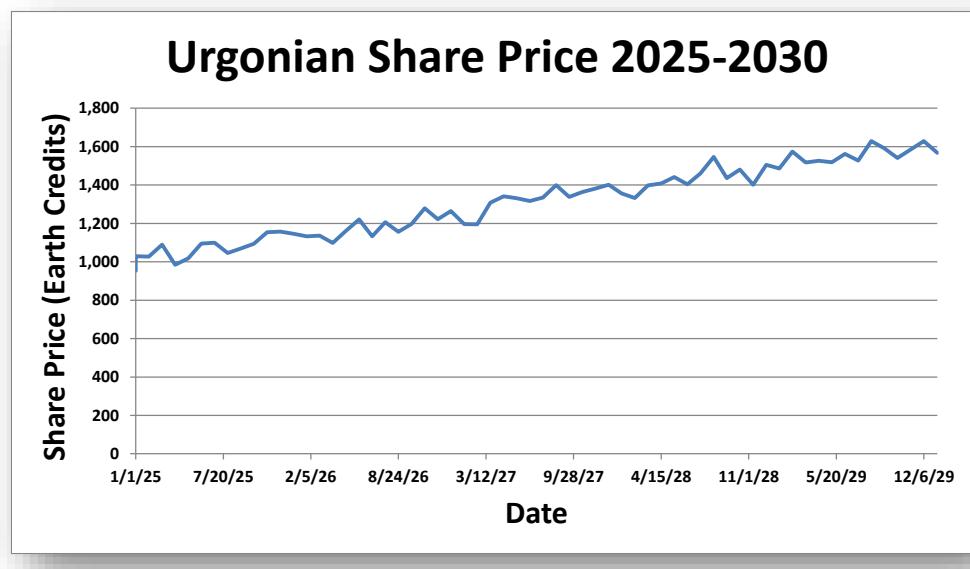
In general, *do not set the intersection of the abscissa and the ordinate (the origin) anywhere but zero* when creating a histogram. Even if you are graphing some other relation not involving frequencies (e.g., price fluctuations in stocks), be wary of arbitrarily cutting off the bottom of the graph and thus potentially misleading viewers into magnifying differences.

INSTANT TEST P 4-23

Experiment with the data in Figure 4-32 to create a *horizontal* bar chart with these values. Change the scale of your new horizontal scale (% responses) to start at 55% and comment on the effect in your chart. Go back to starting the horizontal scale at 0 and compare the impressions created by the two graphs. Why is one better than the other in this case? What would you say if someone told you that there had never been a value lower than 55% in this statistic - would that change your thinking? Why?

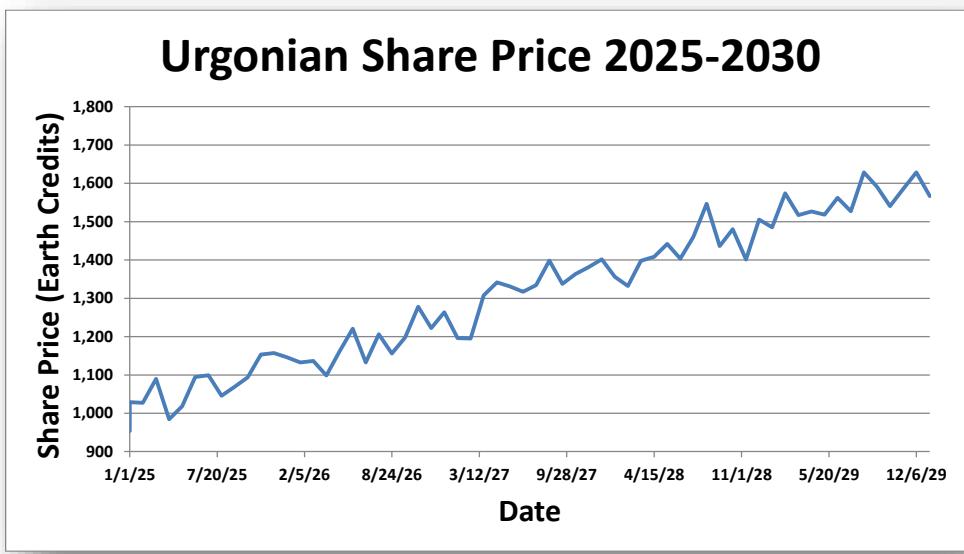
On the other hand, in some applications, it is accepted (despite protests from statisticians) that the origin should reflect the nature of the fluctuations. Figure 4-39 shows prices for a corporation over a five year period. Even without regression analysis or curve fitting, it is obvious that the price has been rising modestly over the period shown.

Figure 4-39. Time series for share prices using 0 as start of ordinate.



However, such time series typically start at some level that reflects the consistent lack of data falling below some arbitrary level. If no one has seen the share price below, say, 900 credits, many financial analysts and journalists will truncate the ordinate as shown in Figure 4-40. Comparing the two graphs, one can see that both the difference between later and earlier prices and the rate of growth in share price seem to be much greater in the graph with the truncated ordinate. This impression might be helpful to dishonest brokers but it could be dangerous for naïve stock traders.

Figure 4-40. Time series for share prices using 900 as start of truncated ordinate (DANGEROUS).



4.14 Selecting Non-Random Sections of a Data Series

One of the most common occurrences of misleading graphs is in published articles where the authors or editors have an axe to grind. By adjusting the origin of the graph and selecting a subset of the available data, they can create whichever impression they wish. For example, examine the carefully crafted graph in Figure 4-41.

Figure 4-41. Misleading graph using selected data and distorted Y-axis (HORRIBLE).



This image is based on the last five stock prices used to create the graph in Figure 4-39 and Figure 4-40 and could be part of an article deliberately trying to give the wrong impression that the share price “has been erratic for some time and is now falling steeply.” Leaving out all the previous data and focusing only on this extract – coupled with setting the Y-axis to give a grossly exaggerated view of the changes – can give whatever impression the writer wants to convey. Whenever you see a grossly truncated ordinate or a small number of data points you should get your skepticism antennae vibrating and look for possible trickery. **And don’t create this kind of monstrosity in your own work!**

INSTANT TEST P 4-25

Find a real time series of data for something you are interested in using resources on the Web or in journals. Prepare a graph for an extended time period. Then take the graph and cut the period to a tiny fraction of the data available, selecting a section that gives the opposite impression of what you see in the overall graph. Show the image of the new image on screen (or in a printout) to some friends and ask them to summarize their impressions about the phenomenon they see displayed in the chart. Then show them the graph for the extended period and ask them what they think about that. Keep a record of the answers and prepare table showing, for each person, their comments on the fragment and their comments on the whole graph. Summarize your findings and post your conclusions in this week’s NUoodle discussion group.

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

5.1 Relative Frequencies, Cumulative Frequencies, and Ogives

We often have to compute the total of the observations in a class and all the classes before it (smaller in an ascending sort or larger in a descending sort). Figure 5-1 shows the *cumulative frequencies* for the ascending sort in **column I**.

The proportion that a frequency represents in relation to the total of the frequencies (the sample size) is called a *relative frequency*. In Figure 5-1, the relative frequencies for the original distribution are shown in **column J**. The relative frequencies for the cumulative distribution are shown in **column K**.

The formulas for computing cumulative and relative frequencies are shown in Figure 5-2, which was generated by choosing the **Formulas | Show Formulas** buttons (green ovals in the figure).

Figure 5-1. Cumulative and relative frequencies.

G	H	I	J	K
Customer Satisfaction (0-100)	Frequency	Cum Freq	Rel Freq	Cum Rel Freq
0	0	0	0%	0%
10	0	0	0%	0%
20	0	0	0%	0%
30	0	0	0%	0%
40	0	0	0%	0%
50	6	6	3%	3%
60	26	32	13%	16%
70	74	106	37%	53%
80	61	167	31%	84%
90	28	195	14%	98%
100	5	200	3%	100%
Totals:	200		100%	

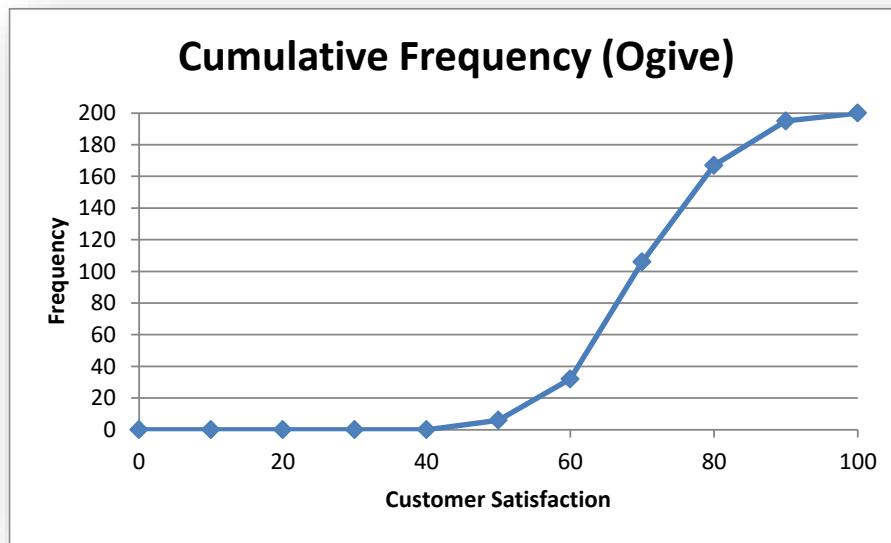
Figure 5-2. Formulas for cumulative and relative frequencies.

G	H	I	J	K
Customer Satisfaction	Frequency	Cum Freq	Rel Freq	Cum Rel Freq
1 0	0	=+H2	=+H2/\$H\$13	=+I2/\$H\$13
2 10	0	=+I2+H3	=+H3/\$H\$13	=+I3/\$H\$13
3 20	0	=+I3+H4	=+H4/\$H\$13	=+I4/\$H\$13
4 30	0	=+I4+H5	=+H5/\$H\$13	=+I5/\$H\$13
5 40	0	=+I5+H6	=+H6/\$H\$13	=+I6/\$H\$13
6 50	6	=+I6+H7	=+H7/\$H\$13	=+I7/\$H\$13
7 60	26	=+I7+H8	=+H8/\$H\$13	=+I8/\$H\$13
8 70	74	=+I8+H9	=+H9/\$H\$13	=+I9/\$H\$13
9 80	61	=+I9+H10	=+H10/\$H\$13	=+I10/\$H\$13
10 90	28	=+I10+H11	=+H11/\$H\$13	=+I11/\$H\$13
11 100	5	=+I11+H12	=+H12/\$H\$13	=+I12/\$H\$13
12 Totals:	=SUM(H2:H12)	=SUM(J2:J12)		

The formula for the first cell in **Column I** for the *Cumulative Frequency* points at the first frequency (**cell H2**). The formula for the second cell in **Column I** (**cell I3**) is the sum of the previous cumulative frequency (**I2**) and the next cell in the *Frequency* column (**H3**). The \$H means that propagating the formula elsewhere maintains a pointer to column H and the \$13 freezes references so that propagating the formula maintains the pointer to row 13.

The line graph for a cumulative frequency is called an *ogive*. Figure 5-3 shows the ogive for the data in Figure 5-1.

Figure 5-3. Ogive for customer satisfaction data.



INSTANT TEST P 5-2

Find some frequency distributions with at least 20 categories in a research paper or statistical report in an area that interests you. Prepare two different frequency distributions based on different bins (e.g., 10 bins vs 20 bins) and create the charts that correspond to each. What are your impressions about using fewer and more categories (bins) in the representation of the frequency data?

5.2 Area under the Curve

One can plot the observed frequencies for the categories defined on the X-axis and examine the area under the curve.

Looking at Figure 5-4, the dark blue line represents the frequency of observations below any particular value of customer satisfaction. The area under the entire curve (shaded pale blue) represents the total number of observations – 200 in this example.

Figure 5-4. Frequency distribution for customer satisfaction scores.



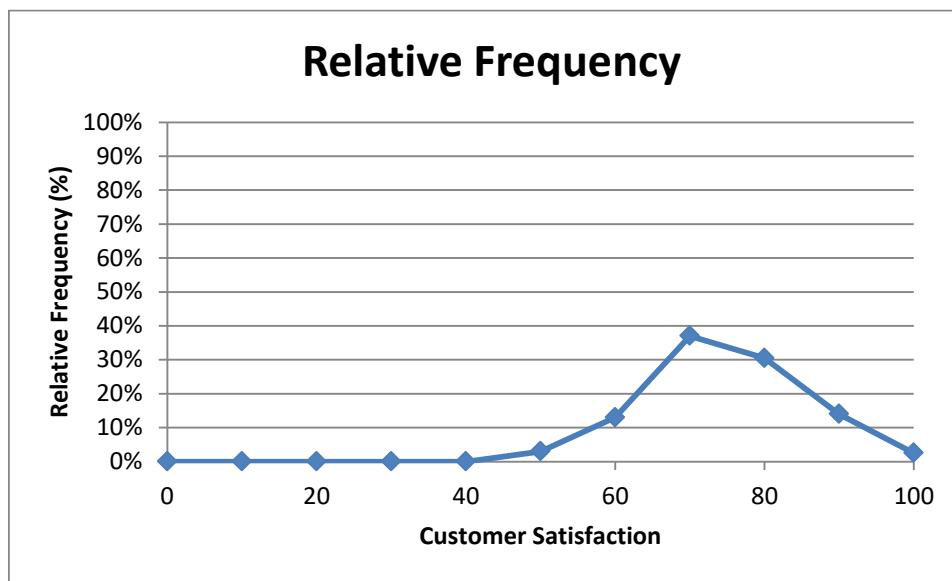
In Figure 5-5, the pale green shaded area represents how many observations were below a specific value of Customer Satisfaction.

Figure 5-5. Areas under curve for observed frequencies.



If one constructs a graph of the frequency distribution with the *relative* frequency data, one ends up with a chart like Figure 5-6.

Figure 5-6. Area under the curve in a relative frequency distribution.



One of the most important principles in using frequency distributions is that the *area under the curve* represents the total proportion of all the observations in all the categories selected. Just as you can add up all the totals for each column in a group of adjacent categories in the histogram, the same principle applies to the frequency distribution.

To repeat, the area under the entire curve represents *100% of the observations*. The area to the *left* of a particular value on the X-axis (e.g., 80) represents the percentage of the observations from zero to just less than the value; e.g., for $X \leq 80$, the area represents the 84% of the total observations (see Column K in Figure 5-1).

5.3 Basic Concepts of Probability Calculations

By definition, an event that is absolutely certain has a probability of one (100%). For example, the probability that a person with a particular disease and specific health attributes will die within the next year is carefully estimated by statisticians working for insurance companies⁵⁵ to help set insurance rates. However, no one needs any sophisticated statistical knowledge to estimate that the likelihood that a person will eventually die some time in the future is 1: that's a known certainty.

Similarly, an impossible event has a probability of zero. Thus, unless one is living in a science fiction or fantasy story, the probability that anyone will be turned inside out during a transporter accident is zero.

If events are mutually incompatible (they can't occur at the same time) and they constitute all possibilities for a given situation, the sum of their individual probabilities is one. For example, a fair cubical die has a probability

$$p_i = 1/6$$

of landing with the top face showing one dot ($i=1$) facing up; indeed $p_i = 1/6$ for all i . The probability that a fair die will land showing a top face of either a 1 or a 2 or a 3 or a 4 or a 5 or a 6 is

$$\sum p_i = 1$$

This makes perfect sense, since the probability of something that is absolutely certain is by definition 1 – and if we exclude weird cases where a die balances on an edge, the only possible faces on the top of a six-sided die are 1, 2, 3, 4, 5, or 6.

If an event i has a probability p_Y then not having the event occur has

$$p_N = 1 - p_Y$$

- Thus the probability that a single throw of a fair die will *not* result with the 2-face upward is $1 - (1/6) = 5/6$. Another way of thinking about that is that there is one out of six ways of satisfying the description of the state and five out of six ways of not satisfying the description of the state.

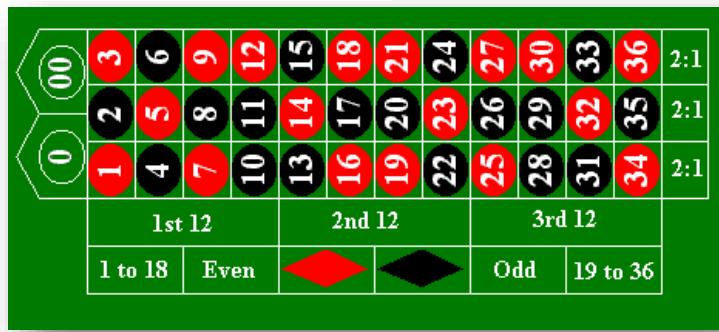
Figure 5-7. US casino roulette wheel.



⁵⁵ Statisticians who work on behalf of insurance companies are called *actuaries*.

- Similarly, since there are 38 slots on a standard roulette US wheel, as shown in Figure 5-7,⁵⁶ the probability that a gambler will win the 36:1 payment for placing a bet on a specific number (e.g., #18) is exactly **1/38**. The probability that the ball will *not* land on slot #18 is exactly 37/38.
- The probability that a roulette player will win the 2:1 payout for having a bet on the red box on the board when the ball lands in a red slot is exactly **18/38** and the probability that the ball will *not* land on a red slot is therefore $1 - (18/38) = 20/38$.

Figure 5-8. US casino roulette board.



If events are *independent* of each other (not influenced by each other), the the probability that two events p_1 and p_2 will occur at once (or in sequence) is

$$p_1 * p_2$$

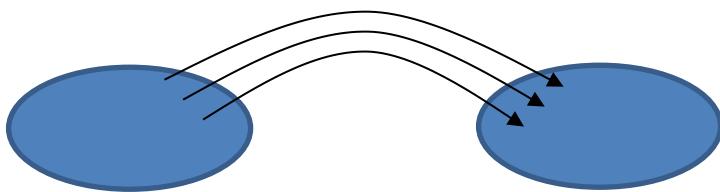
- For example, if you buy a lottery ticket with a 1 in 100,000 chance of winning \$10,000, the chance of winning \$10,000 is therefore 1/100,000. If you buy two of the same kind of lottery tickets, the chance of winning \$10,000 on *both* of the tickets is $(1/100,000) * (1/100,000) = 1/10,000,000,000$ or simply $1e-5 * 1e-5 = 1e-10$.
- The probability of *losing* on *both* lottery tickets is $(1 - 1e-5) * (1 - 1e-5) = 0.99999 * 0.99999 = 0.99998$. The probability of winning on *at least one* ticket is $1 - 0.99998 = 0.00002$.

This kind of reasoning is especially useful in calculating failure rates for complex systems. In information technology, a useful example of the probability-of-failure calculations is *Redundant Arrays of Independent Disks* – specifically RAID 1 and RAID 0 disk drives.

Here are the basics about these two configurations:

- RAID 1 (redundancy) improves resistance to disk failure (i.e., provides fault tolerance) by making bit-for-bit copies of data from a main drive to one or more mirror drives. If the main drive fails, the mirror drive(s) continue(s) to provide for I/O while the defective drive is replaced. Once the new, blank drive is in place, array management software can rebuild the image on the new drive. The frequency of mirroring updates can be defined through the management software to minimize performance degradation. As long as at least one of the disks is working, the array works.

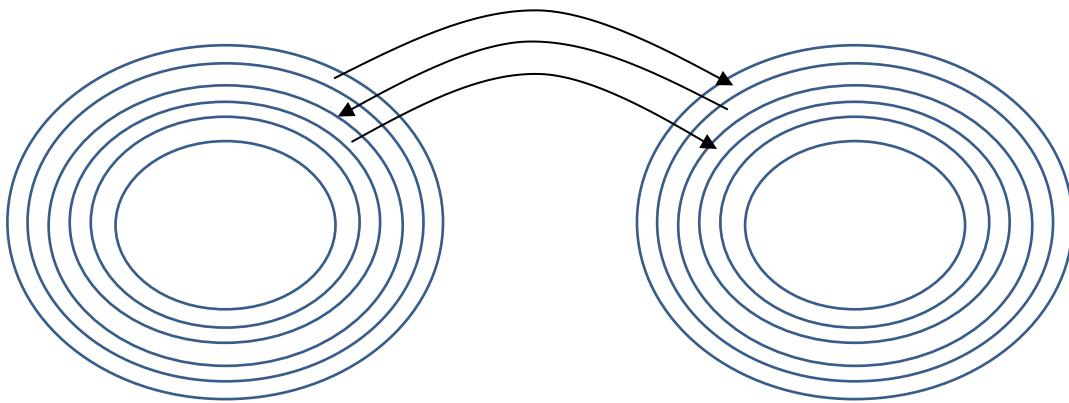
Figure 5-9. Raid 1 array with 2 disks showing writing from primary disk to secondary disk.



⁵⁶ There are 18 black slots and 18 red slots – all of which are involved in paying out money to the gambler depending on the bets – and two house slots (0 and 00) that result in having the house take all of the bets on the table without paying anything out.

- RAID 0 (speed) improves performance by *striping*, in which data are written alternately to cylinders of two or more disk drives. With multiple disk heads reading and writing data concurrently, input/output (I/O) performance improves noticeably. All the disk drives in the array must work for the array to work.

Figure 5-10. RAID 0 array showing writing to cylinders in alternate disks.



Now let's think through the probability of failure for each of these RAID configurations.

- Let the probability of failure of any one disk in a specified period be p (e.g., $1/100$ per year = 0.01).
- For a RAID 1 (redundancy) array with n independent and identical disks, the probability that all n disks will fail is

$$P\{\text{all } n \text{ drives fail}\} = p^n$$

For example, with $p = 0.01/\text{year}$ and two disks in the RAID 1 array, the chances that both disks will fail at the same time is only 0.01^2 or **0.0001** (one failure in a ten thousand arrays).

- For a RAID 0 (speed) array with n interleaved independent and identical disks, every disk must function at all times for the array to work. So first we compute the probability that all the drives will work.

$$P\{\text{all } n \text{ drives work}\} = (1 - p)^n$$

- For example, using the same figures as in the previous bullet, we compute that the chance of having both drives work throughout the year is $0.99^2 = 0.9801$.
- But then the chance that *at least one of the drives* will not work must be

$$P\{\text{at least one drive fails}\} = 1 - (1 - p)^n$$

and therefore the example gives a failure probability for the RAID 0 of $1 - 0.9801 = 0.0199$ – almost double the probability of a single-disk failure.

- If there were 10 disks in the RAID 0, the probability of RAID failure using the same figures as the examples above would be

$$1 - (1 - 0.01)^{10} = 1 - 0.99^{10} = 1 - 0.9043821 = 0.0956$$

or almost 10 times the single-disk failure.

The same kind of reasoning applies even if the probabilities of elements in a system are different. One then uses the following formulae:

- For redundant systems which work if *any* of the n components ($p_1, p_2, p_3 \dots p_n$) in the calculation work, so that we need the probability that *all the components will fail*,

$$P\{\text{system failure}\} = p_1 * p_2 * p_3 * \dots * p_n$$

which is more economically represented using the capital pi symbol (Π) for multiplication (much like the capital sigma (Σ) symbol for addition),

$$P\{\text{system failure}\} = \prod p_i$$

- Similarly, for a redundant system,

$$\begin{aligned} P\{\text{system failure}\} &= \\ 1 - [(1 - p^1) * (1 - p^2) * (1 - p^3) * \dots * (1 - p^n)] &= \\ 1 - \prod (1 - p_i) \end{aligned}$$

INSTANT TEST P 5-8

Without referring to this text or to notes, explain the reasoning for how to calculate the probability that a two-disk RAID 1 (redundancy) array will fail within a year if the probability that a disk will fail in a year is known. Then explain how to calculate the probability of failure within a year for a RAID 0 (speed) array with three disks.

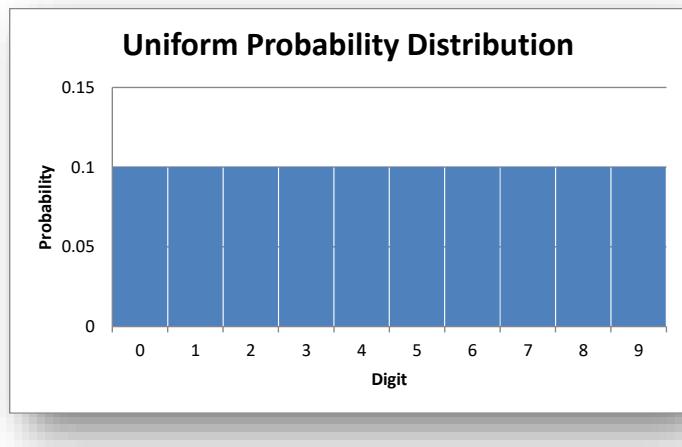
5.4 The Uniform Probability Distribution

As a simple example, suppose we count the number of each of the 10 digits (0, 1, 2.... 9) in an long series of numbers generated by a good random-number generator such as Excel's =RAND() or =RANDBETWEEN(lower, upper) functions. In the long run, we would expect 10% of all the digits to be 0, 10% to be 1, and so on. Figure 5-11 shows the *uniform probability distribution* corresponding to this thought-experiment. Each digit has a 10% (0.1) probability of being observed in a random sequence. We call this probability the *probability density function* and describe it as

$$P(x) = 0.1$$

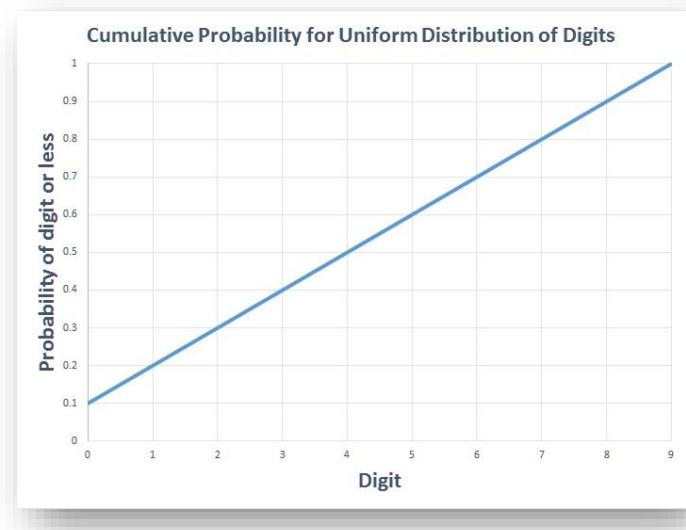
for all 10 values of x, the random deviate. Figure 5-11 shows the uniform distribution graphically.

Figure 5-11. Probability distribution for digits in random numerical data.



Now consider how many of the digits will be 3 or lower (3, 2, 1, 0): clearly there should be 40% (0.4) that correspond to that selection criterion. Accordingly, the *surface area under the curve* should be 40% – and that matches what we see in

Figure 5-12. Cumulative probability distribution for digits in random numerical data.



The area under the “curve” (in this case it’s a straight line across the top of the rectangles) is simply

$$P(x \leq n) = 0.1(n + 1)$$

For example, the probability that a randomly selected digit will be 6 or less is $0.1(6+1) = 0.7$, which matches what we can see on the graph. We call this function the *cumulative probability* function, often called *alpha* (α).

Finally, if we want to know how large the digit x_α is that includes $\alpha=0.8$ of the observations at or below its value, we can see on the graph that $x = 7$ satisfies the requirement. The function is notated as

$$x_\alpha = 10\alpha - 1$$

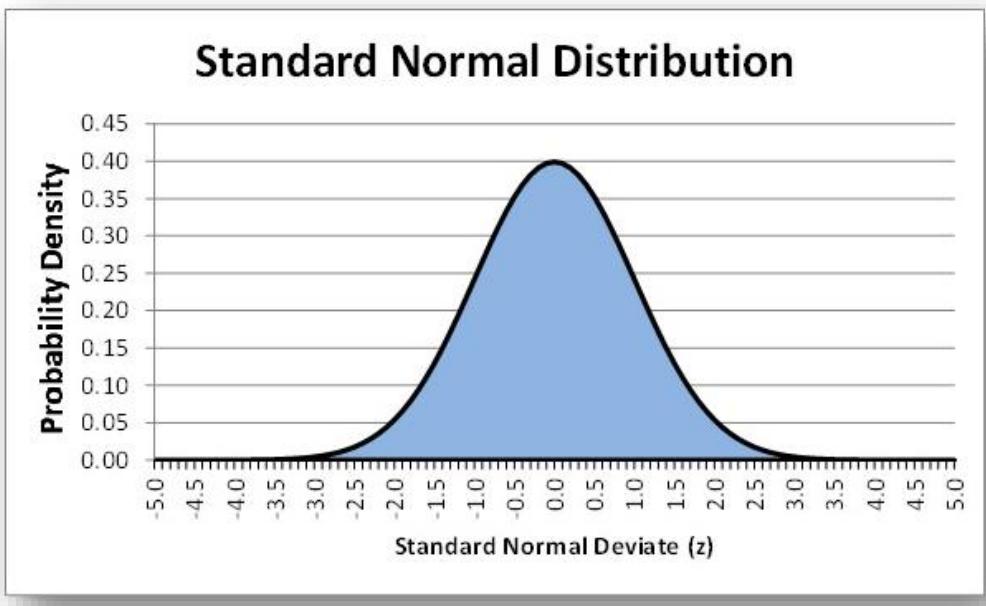
and is called the *inverse function* for the uniform distribution. Other distributions have their own formulas for calculating the probability density, the cumulative probability, and the inverse function. More important for our use, they are also computed for us automatically in EXCEL and in other statistical packages.

5.5 The Normal Probability Distribution

There are many theoretical probability distributions that we use in statistics. For example, the standard Normal distribution describes how observations vary when there are many factors that come together to generate the observation. A familiar example is peoples' weight: many factors contribute to the weight of an individual, and the *bell-curve*⁵⁷ shown in Figure 5-13 illustrates the kind of variation one would expect by computing the *standard Normal deviate*, z , based on the average (arithmetic mean) and on how much variability (measured by the standard deviation) there is in the population. We'll be studying these descriptive statistics in detail later.

In Figure 5-13, the dark bell-shaped curved *line* is the *probability density* function; this is a function that allows us to compute the probability that a deviate occurs in any given interval. Probability density functions are not probabilities: they often exceed the value 1.⁵⁸

Figure 5-13. Standard Normal curve illustrating probability distribution.



The blue fill under the curve symbolizes the *area under the curve*; as always, the total is defined as 1 (100%) because all possible values of z are considered.

A probability of 1 means *certainty*; it is certain that the values of z must lie between $-\infty$ and $+\infty$; however, as you can see from the graph, the probability density function falls to infinitesimal levels around $z = -4$ and $z = +4$.

Half of this distribution lies at or below the mean and half falls above the mean.

⁵⁷ Also called a *Gaussian* distribution in honor of Carl Friedrich Gauss, who lived from 1777 to 1855 in Germany and who contributed greatly to probability theory, statistics, and many other aspects of science and mathematics. See (Gray 2012).

⁵⁸ For exhaustive details of the probability density function, see (Nykamp nd).

To find the exact values relating to the Normal distribution, we can use the Normal distribution functions in EXCEL, which several functions related to the Normal distribution, as shown in Figure 5-14. One can access this menu by typing =NORM in a cell.

These functions are explained in the **HELP** facility for EXCEL 2010, as shown in . The versions of the functions without the period separators are included in EXCEL for compatibility with older versions of EXCEL.

Figure 5-14. Normal distribution functions in Excel 2010.

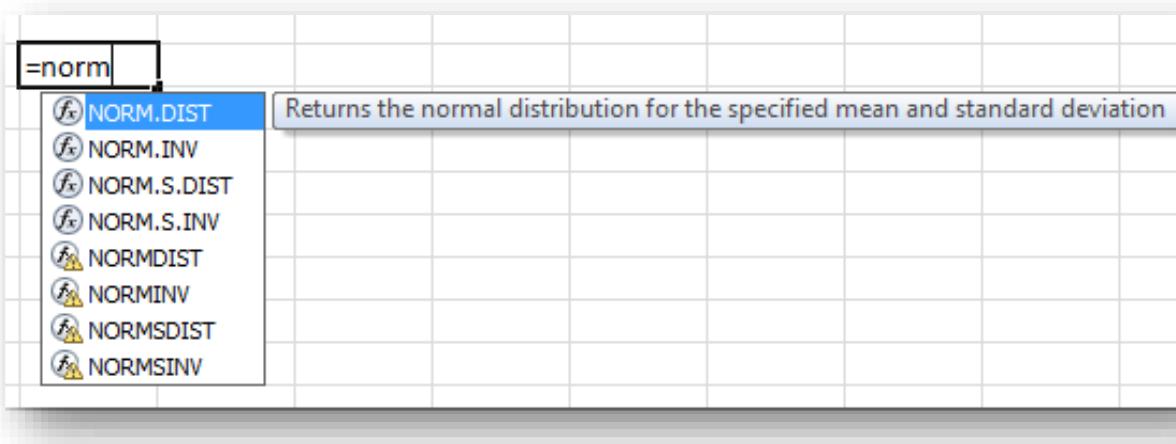
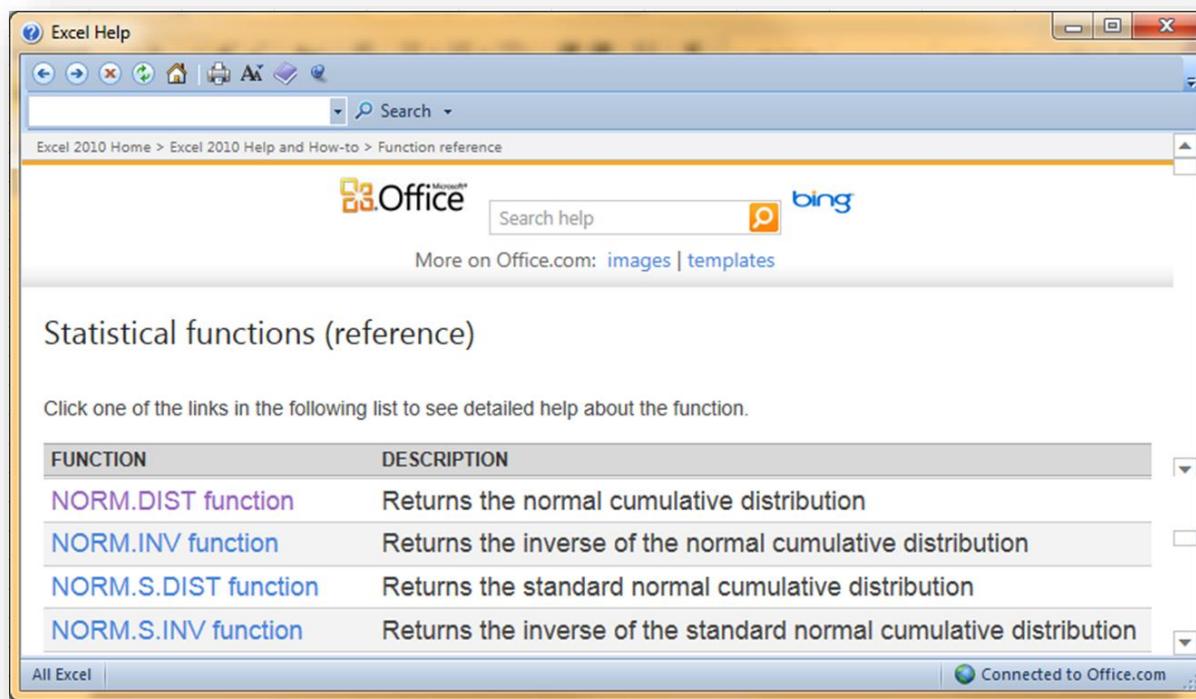


Figure 5-15. Excel 2010 HELP for Normal distribution functions.



Excel Help

Search

Excel 2010 Home > Excel 2010 Help and How-to > Function reference

Office bing

Search help

More on Office.com: images | templates

Statistical functions (reference)

Click one of the links in the following list to see detailed help about the function.

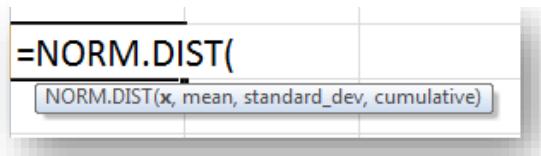
FUNCTION	DESCRIPTION
NORM.DIST function	Returns the normal cumulative distribution
NORM.INV function	Returns the inverse of the normal cumulative distribution
NORM.S.DIST function	Returns the standard normal cumulative distribution
NORM.S.INV function	Returns the inverse of the standard normal cumulative distribution

All Excel Connected to Office.com

5.6 Area under the Curve for Any Normal Distribution

In applied statistics, you will constantly have to find the area under a theoretical probability distribution curve. The area of interest is generally symbolized by lower-case alpha (α). For the Normal distribution, the =NORM.DIST (=NORMDIST in the older versions of EXCEL) function returns (calculates) the area of the curve to the *left* of a specified Normal deviate. For example, typing =NORM.DIST (or left-clicking on that function in the menu shown in Figure 5-14 instantly brings up the *parameters* required for the function to work, as shown in Figure 5-16.

Figure 5-16. Pop-up reference to parameters for =NORM.DIST function in Excel 2010.



The first parameter, x , is the value of interest; **mean** is the average of the distribution, **standard_dev** is the standard deviation (σ , pronounced *sigma*, which we will study in detail later) and **cumulative** is 0 for the *probability density function* and 1 for the *cumulative probability* (area of the curve to the left of) for x .

Intuitively, if we have a Normal distribution such as the IQ distribution, which theoretically has mean IQ = 100 and standard deviation = 15, half the values will lie at or below the mean. So calculating

$$=NORM.DIST(100,100,15,1)$$

should give us $\alpha = 0.5$. And sure enough, it does, as shown in Figure 5-17.

Figure 5-17. Calculating a cumulative probability value using NORM.DIST in Excel 2010.

<i>f</i> x	=NORM.DIST(100,100,15,1)
C	D
0.5	

Given the appearance of the bell curve and the logic of probabilities for mutually exclusive cases, the probability that a random x will be greater than the mean for this example is given by

$$P(x > 100) = 1 - P(x \leq 0) = 1 - 0.5 = 0.5$$

INSTANT TEST P 13

Practice using the =NORM.DIST function with different values to ensure that you become familiar with it. As a check on your work you can remember that a Normal distribution with mean = 0 and standard_dev = 1 has 0.025 to the left of $x = -1.96$ and 0.975 to the left of $x = +1.96$. Play around with different values of the variables. You may want to point to cells for the variables you are changing rather than having to modify constants in the formula itself. Be sure to experiment with the cumulative parameter to see for yourself what it does.

5.7 Area Under the Curve for the Standard Normal Distribution

The standard Normal deviate z is

$$z = (Y - \mu)/\sigma$$

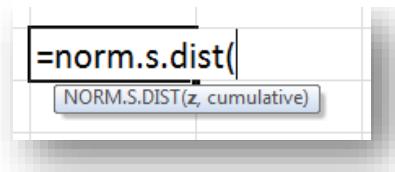
where

- Y is a value for the abscissa
- μ is the mean of the population
- σ is the parametric standard deviation of the population.

The `=NORM.S.DIST` function generates the probability that random observation from a Normal distribution will generate a standardized Normal deviate at or below the indicated value z as shown in Figure 5-19.⁵⁹

The second parameter, `cumulative`, is 1 for the cumulative distribution and 0 for the (rarely-used) probability density function.

Figure 5-19. `=NORM.S.DIST` function in Excel 2010.



For example, the area under the curve at and to the left of $z = 2$ is 0.97725, as shown in Figure 5-18. In other words, the probability of observing a $z \leq 2$ by chance in a standardized Normal distribution is 97.725%.

The `=NORM.S.INV` function provides the *critical value* of z corresponding to the α for the left tail of the distribution. How big must z be to demarcate a given left-tail α ? Figure 5-20 shows that the value of $z = 1.96$ demarcates a left-tail area under the standard Normal curve of 0.975. Thus the area on the right must be $1 - 0.975 = 0.025$.

Because the Normal curves are symmetrical, this finding also implies that 0.025 of the curve lies to the *left* of -1.96 . And indeed the function `=NORM.S.INV(0.025)` does generate -1.9600 (not shown).

Figure 5-18. Calculation of probability that $z \leq 2$ in Excel 2010.

f_x	<code>=NORM.S.DIST(2,1)</code>
J	0.97725

Figure 5-20. Finding the critical value for a left-tail area in the standard Normal curve using an Excel 2010 function.

f_x	<code>=NORM.S.INV(0.975)</code>
D	1.9600

⁵⁹ By accident, the function name was typed in lowercase before this screenshot was taken. Uppercase and lowercase don't matter when typing function names in Excel.

5.8 Using EXCEL Functions for Areas Under Other Probability Distribution Curves

In addition to the Normal distribution, other statistical probability distributions that you will use extensively are the

- Chi-square (χ^2): important for *goodness-of-fit tests* (seeing if sampled data fits a theoretical expected distribution) and *tests of independence* (seeing if the underlying proportions of different classes are the same in different samples);
- F: for comparing the underlying *variances* (measure of variability) of two samples; particularly important in the valuable *analysis-of-variance (ANOVA)* methods;
- Student's t: for comparing the underlying statistics from different samples to see if they are the same; also for computing *confidence limits* (estimates of what the statistics could actually be given the sample values) to a wide variety of statistics.

The basic steps for learning about a statistical function in any computational package are to

- Find the functions by using a reasonable identifier – in EXCEL, one can start by typing out =name (e.g., =norm) where name represents the name of the distribution;
- Learn what the parameters are (in EXCEL, check the pop-ups and use HELP);
- Try computing something for which you know the answer to check that you're doing it right (e.g., use statistical tables for known values and see if you can duplicate the results);
- You can also compute complementary values to check that you're using the right function – e.g., calculating =norm.s.dist(0.05) = -1.644854 and =norm.s.dist(0.95) = 1.644854.

Some commonly used terminology for statistical functions includes

- *Left-tailed probability*: likelihood of observing a value *smaller than or equal to* the test value if the assumptions of the statistical function are true;
- *Right-tailed probability*: likelihood of observing a value *larger* (or sometimes *larger than or equal to*) than the test value if the assumptions of the statistical function are true.

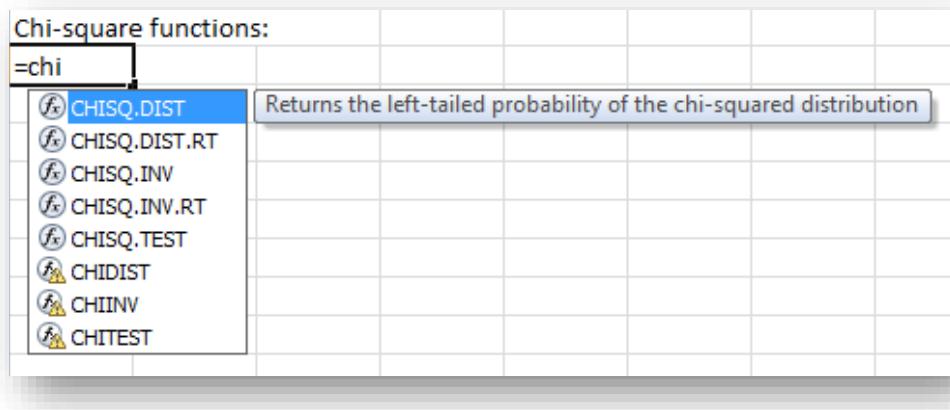
In general, there is little or no difference in practice between the area of a probability distribution to the *left of a value* ($< x$) and the area to the *left of or equal to a value* ($\leq x$).

5.9 Chi-Square Distribution

Typing =chi into a cell in EXCEL 2010 brings up a menu of choices, as shown in Figure 5-21.

Clicking on any one of the functions generates a list of its parameters, as shown in Figure 5-22. The middle

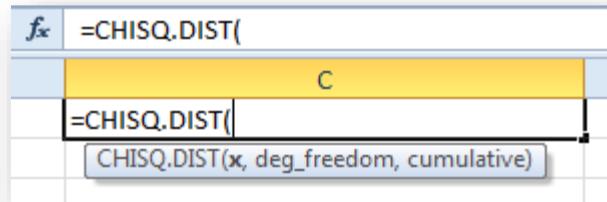
Figure 5-21. Menu of chi-square functions in Excel 2010.



term, *deg_freedom*, refers to *degrees of freedom* which are a function of the particular application of the chi-square.

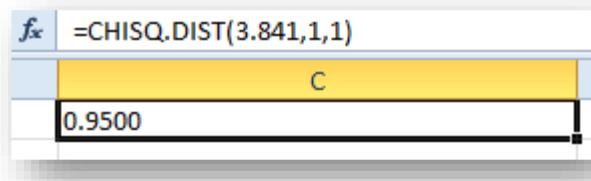
To check that you are using the =CHISQ.DIST function correctly, you can remember that 95% of the curve

Figure 5-22. Left-tail chi-square probability function in Excel 2010.



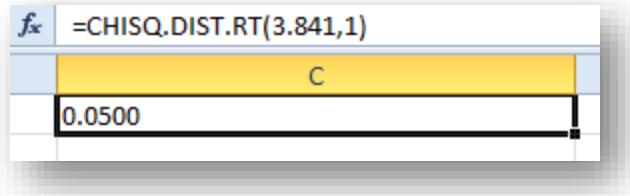
lies to the left of 3.841 when the distribution has 1 *degree of freedom*, as shown in Figure 5-23.

Figure 5-23. Left-tail chi-square probability function in Excel 2010.



That value also implies that 5% of the curve (1-0.95) lies to the *right* of 3.841 for 1 degree of freedom, as confirmed in Figure 5-24 using the EXCEL 2010 function =CHISQ.DIST.RT.

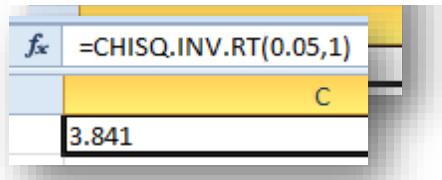
Figure 5-24. Right-tail chi-square probability function in Excel 2010.



The inverse functions =CHISQ.INV and =CHISQ.INV.RT produce the critical values for left-tailed and right-tailed probabilities, respectively. Thus for a left-tailed $\alpha = 0.95$ with one degree of freedom, the critical value is 3.841, as shown in Figure 5-25.

Figure 5-25. Critical value for left-tail = 0.95 in chi-square distribution using Excel 2010 function.

Figure 5-26. Critical value for right tail = 0.05 in chi-square distribution using Excel 2010 function.



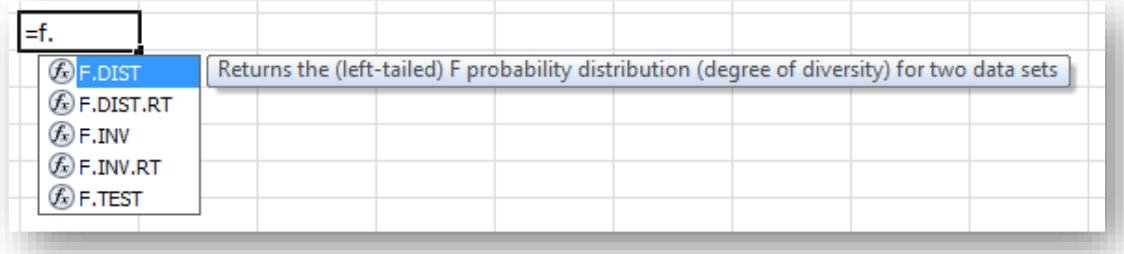
That finding also means that the right tail must be 0.05, and that's confirmed also, as shown in Figure 5-26.

We'll be looking at the =CHISQ.TEST function later in the course, where it plays an important role in testing *frequency distributions* for *categorical* data against theoretical distributions – tests of *goodness of fit* and of *independence*. It actually calculates a value of a sample chi-square that can be tested against the theoretical distribution to see how likely it is that such a large value could occur by chance alone if the assumptions of the test were correct. The =CHITEST function, preserved in EXCEL for compatibility with older versions of EXCEL, generates the same result.

5.10 F Distribution

The F-distribution functions in EXCEL 2010 are shown in Figure 5-27. F-tests are critically important in analysis of variance (ANOVA) methods widely used in applied statistics.

Figure 5-27. F-distribution functions in Excel 2010.



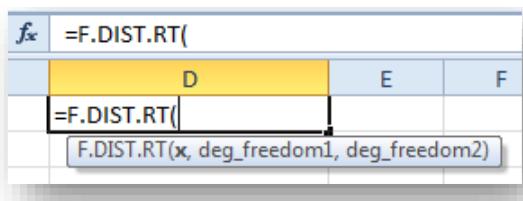
The HELP facility documents the functions, as shown in Figure 5-28.

Figure 5-28. HELP for F-distribution functions in Excel 2010.

F.DIST function	Returns the F probability distribution
F.DIST.RT function	Returns the F probability distribution
F.INV function	Returns the inverse of the F probability distribution
F.INV.RT function	Returns the inverse of the F probability distribution
F.TEST function	Returns the result of an F-test

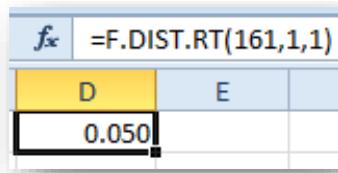
Each function is documented in the HELP facility by clicking on the link in Figure 5-28. In addition, starting to type the function in an EXCEL cell brings up a pop-up reminding the user of the parameters required, as shown in Figure 5-29.

Figure 5-29. Pop-up menu in Excel 2010 for right-tail area under the curve for F distribution.



For reference, you can note that the value F = 161 cuts off 5% of the distribution curve for 1 and 1 degrees of freedom, as shown in Figure 5-30.

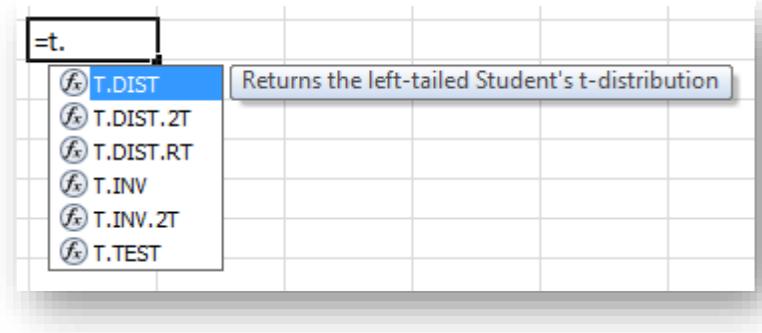
Figure 5-30. Critical value for right-tailed probability of 0.05 in F distribution with 1 & 1 df in Excel 2010.



5.11 Student's-t Distribution

In EXCEL 2010, typing =T. (or =t.) brings up the menu of *Student's t*⁶⁰ functions as shown in Figure 5-31. The =T.DIST function generates the left tail of the distribution. For example, $t = 12.706$ for a distribution

Figure 5-31. Student's t functions in Excel 2010.



with 1 degree of freedom cuts off 97.5% to the left of that value, as shown in Figure 5-32.

The symbol .2T in the name of the function indicates a *two-tailed probability*, as shown in Figure 5-33.

The symbol .RT in the function name indicates a right-tailed probability, as shown in Figure 5-34.

Figure 5-32. Left-tail probability for t distribution in Excel 2010.

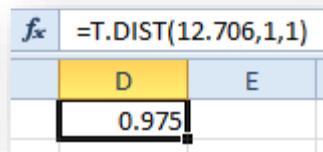


Figure 5-33. Two-tailed probability for t distribution in Excel 2010.

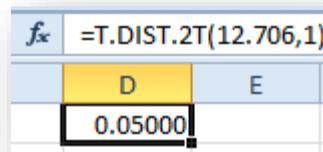
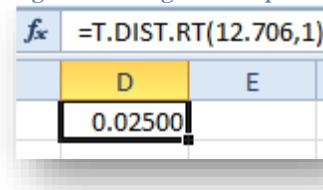


Figure 5-34. Right-tailed probability for t distribution in Excel 2010.



⁶⁰ Described in English in 1908 by William S. Gosset, who published under the pseudonym *Student*. See (Encyclopaedia Britannica 2012).

6 Descriptive Statistics

6.1 Summarizing Groups of Data using EXCEL Descriptive Statistics

Listing raw data and even sorted data becomes confusing when we deal with many observations. We lose track of overall tendencies and patterns in the mass of detail. Statisticians have developed a number of useful methods for summarizing groups of data in terms of approximately what most of them are like (the *central tendency*) and how much variation there is in the data set (*dispersion*). When more than one variable is involved, we may want to show relations among those variables such as breakdowns in the numbers falling into different classes (*cross-tabulations* or *contingency tables*), measures of how predictable one variable is in terms of another (*correlation*) and measures of how to predict the numerical value of one variable given the value of the other (*regression*).

Descriptive statistics have two forms:

- *Point estimates*, which indicate the most likely value of the underlying phenomenon; e.g., “the mean of this sample is 28.4”
- *Interval estimates*, which provide a range of values with a probability of being correct; e.g., “the probability of being correct is 95% in stating that the mean of the population from which this sample was drawn is between 25.9 and 30.9”

The EXCEL Data | Data Analysis sequence (Figure 6-1) brings up a useful Data Analysis tool called Descriptive Statistics shown in Figure 6-2.

Figure 6-1. Excel 2010 Data Analysis menu.

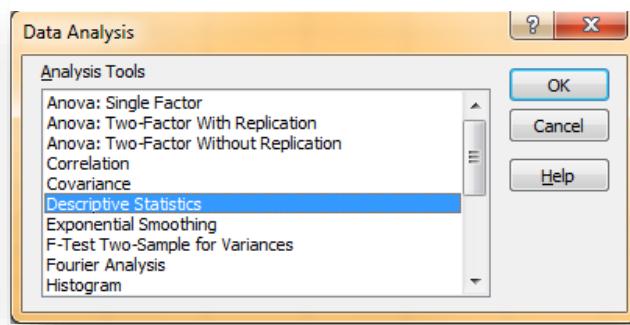
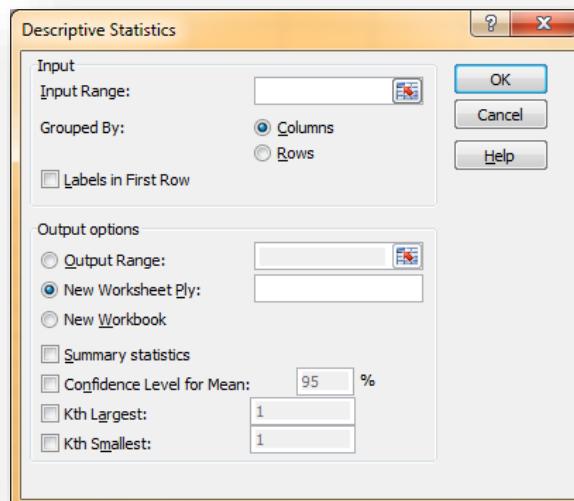


Figure 6-2. Excel 2010 Descriptive Statistics pop-up.



For example, a series of intrusion-detection-system log files from a network documents the numbers of network-penetration attempts (Network Attacks) per day for a year. Figure 6-3 shows the **Descriptive Statistics** data analysis tool filled out to accept columnar data with the first line as a heading or label; it is also set to locate the 5th largest and the 5th smallest entries as an illustration.⁶¹ The 95% **Confidence Level for Mean** setting generates the amount that must be subtracted and added to the mean to compute the 95% confidence limits for the mean., discussed in detail later in the course.

Applying the **Descriptive Statistics** analysis tool with these settings produces the summary shown in Figure 6-4. In the sections following, we'll discuss each of these results.

The results of the **Descriptive Statistics** tool *do not change dynamically*: if the data are modified, you have to apply the tool again to the new data. Unlike functions, which instantly show the new results if the input data are changed, the results produced by any of the Data Analysis tools are static and must be recalculated explicitly to conform to the new data.

Figure 6-3. Descriptive Statistics tool in Excel 2010.

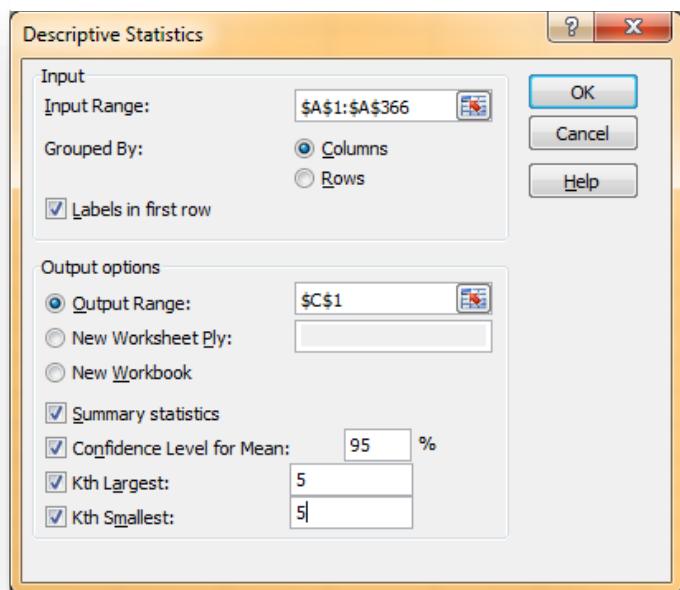


Figure 6-4. Descriptive Statistics results.

Network Attacks	
Mean	2844.425
Standard Error	25.64544
Median	2859
Mode	2872
Standard Deviation	489.9554
Sample Variance	240056.3
Kurtosis	-0.14605
Skewness	-0.03163
Range	2552
Minimum	1632
Maximum	4184
Sum	1038215
Count	365
Largest(5)	3936
Smallest(5)	1692
Confidence Level(95.0%)	50.43182

INSTANT TEST P 6-2

Generate some Normally distributed random data using =int(norm.inv(rand(),mean,std-dev)). Explain exactly what each part of this function does by showing it to a buddy who doesn't know Excel.

Practice using the Descriptive Statistics tool on your data.

Notice that the data change constantly. Apply the Descriptive Statistics tool again and place the results below the original output so you can compare them.

Observe that the values are different (except the Count).

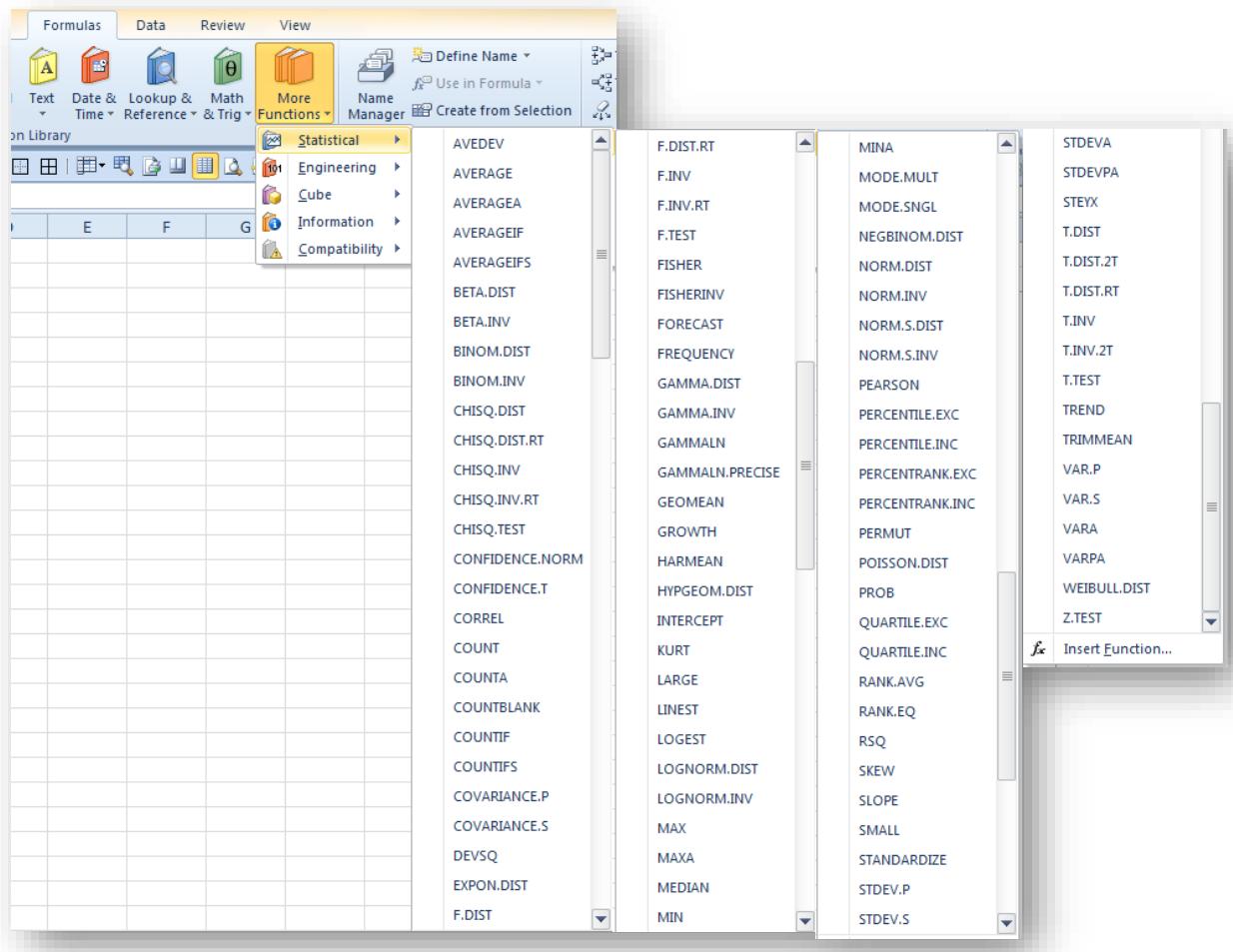
⁶¹ For example, in exploratory data analysis, one might want to examine the most extreme five data on each side of the median to see if they were outliers to be excluded from further analysis.

6.2 Computing Descriptive Statistics using Functions in EXCEL

In the next section, we'll examine a number of descriptive statistics, including the ones listed in Figure 6-4. Before that, though, it's useful to know that EXCEL can also compute individual descriptive statistics using functions. For example, Figure 6-5 shows all the individual statistical functions available in EXCEL 2010 through the **Formulas | More Functions | Statistical** sequence.

As always, every function is fully described in the EXCEL **HELP** facility. The functions with periods in their name are defined in EXCEL 2010 and, for the most part, correspond to the equivalent functions in earlier versions of EXCEL.

Figure 6-5. Statistical functions in Excel 2010.



INSTANT TEST P 6-3

To ensure that you remember how to access these functions quickly any time you need them, practice bringing up the function lists for all the different functional areas available in the Function Library portion of the Formulas menu.

For a few functions that you recognize and for some you don't, bring up the Help facility to see the style of documentation available that way.

6.3 Statistics of Location

At an intuitive level, we routinely express our impressions about representative values of observations. Even without calculations, we might say, “I usually score around 85% on these exams” or “Most of the survey respondents were around 50 years of age” or “The stock seems to be selling for around \$1200 these days.” There are three important measures of what statisticians call the *central tendency*: an average (the arithmetic mean is one of several types of averages), the middle of a ranked list of observations (the *median*), and the most frequent class or classes of a frequency distribution (the *mode* or *modes*). In addition, we often refer to groups within a distribution by a variety of *quantiles* such as *quartiles* (quarters of the distribution), *quintiles* (fifths), *deciles* (tenths) and *percentiles* (hundredths).⁶² These latter statistics include attributes of statistics of dispersion as well as of central tendency.

6.4 Arithmetic Mean (“Average”)

The most common measure of central tendency is the *arithmetic mean*, or just *mean* or *average*.⁶³ We add up the numerical values of the observations (Y) and divide by the number of observations (n):

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{i=n} Y_i \quad \text{or more simply} \quad \bar{Y} = \frac{\sum Y}{n}$$

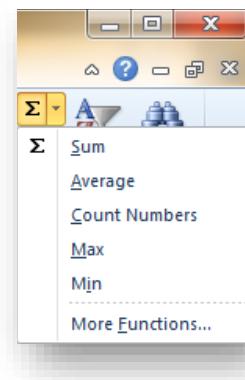
In EXCEL, we use the =AVERAGE(data) function where data is the set of cells containing the observations, expressed either as a comma-delimited list of cell addresses or as a range of addresses. Figure 6-7 shows an average of three values computed with the =AVERAGE(data) function.

Figure 6-7. Excel AVERAGE function.

	A	B	C
1	15		
2	22		
3	43		
4	26.66667	=AVERAGE(A1:A3)	

For convenience, several frequently used functions are accessible in a drop-down menu on the main EXCEL bar, as shown in Figure 6-6. The functions are accessible when a range of data has been selected.

Figure 6-6. Drop-down menu in Excel 2010.



⁶² The percentile is occasionally called a *centile*.

⁶³ The arithmetic mean is distinguished from the geometric mean, which is used in special circumstances such as phenomena that become more variable the larger they get. The geometric mean is the nth root of the product of all the data. E.g., for data 3, 4, 5, the geometric mean is $(3*4*5)^{(1/3)} = 60^{(1/3)} = 3.915$. In Excel the =geomean() function computes the geometric mean of a range.

6.5 Calculating an Arithmetic Mean from a Frequency Distribution

Sometimes we are given a summary table showing means for several groups; for example, suppose three different soap brands have been tested in a particular market. Cruddo was introduced first, then Gloppu and finally Flushy. Figure 6-8 shows the average sales figures for each of three brands of soap along with the total number of months of observations available.

Figure 6-8. Monthly sales figures and weighted average.

Soap	Average Monthly Sales	Number of Months in Study	Total sales
Cruddo	\$38,547	20	\$770,940
Flushy	\$37,593	12	\$451,116
Gloppu	\$27,379	16	\$438,064
Totals		48	\$1,660,120
	<i>Weighted average:</i>		\$34,586

It wouldn't make sense just to add up the average monthly sales and divide by three; that number (\$34,506) doesn't take into account the different number of months available for each soap. We simply multiply the average monthly sales by the number of months (the *weight*) to compute the total sales for each soap (the *weighted totals*) and divide the sum of the weighted totals by the total number of months (the total of the weights) to compute the *weighted average* as shown in Figure 6-8. The weighted average, \$34,586, accurately reflects the central tendency of the combined data.

Figure 6-9 shows the formulas used to compute the weighted average in this example.

Figure 6-9. Formulas used in previous figure.

Soap	Average Monthly Sales	Number of Months in Study	Total sales
Cruddo	38547	20	770940
Flushy	37593	12	451116
Gloppu	27379	16	438064
Totals		=SUM(C2:C4)	=SUM(D2:D4)
	<i>Weighted average:</i>		=D5/C5

INSTANT TEST P 6-5

Duplicate the example yourself in Excel. Compute the erroneous average of the Average Monthly Sales using the appropriate Excel function to see for yourself that it's wrong.

6.6 Effect of Outliers on Arithmetic Mean

The arithmetic mean has a serious problem, though: it is greatly influenced by exceptionally large values – what we call *outliers*.

Here's an imaginary list of five household annual incomes (Figure 6-10).

Figure 6-10. Household incomes showing a wealthy outlier.

Household Income	\$ 18,243	\$ 20,234	\$ 38,481	\$ 42,945	\$ 8,343,591
------------------	-----------	-----------	-----------	-----------	--------------

The mean income for these five households is \$1,692,699. Does that value seem representative to you? The average for the four non-wealthy incomes is \$29,976: about 2% of the computed (distorted) average including the outlier. Another way of looking at the problem is that the distorted mean is 56 times larger than the mean of the lower four incomes.

If you think about how the average is computed, it makes sense that unusually large outliers can grossly distort the meaning of the arithmetic mean. One way of reducing such an effect is to exclude the outliers from the computation of the average. For example, one can set an exclusion rule in advance that leaves out the largest and the smallest value in a data collection before performing statistical analysis on the data. However, these rules can be problematic and should be carefully discussed and thought through before applying them to any scientific or professional study.

Do unusually small values also distort the arithmetic mean? In Figure 6-11 we see another made-up sample, this time with mostly millionaires and one poor family.

The average of the *five* household incomes is \$6,308,938. Does the poor family unduly influence the

Figure 6-11. Four wealthy families & one poor outlier.

Household Income	\$ 18,243	\$ 7,292,153	\$ 7,753,481	\$ 8,137,222	\$ 8,343,591
------------------	-----------	--------------	--------------	--------------	--------------

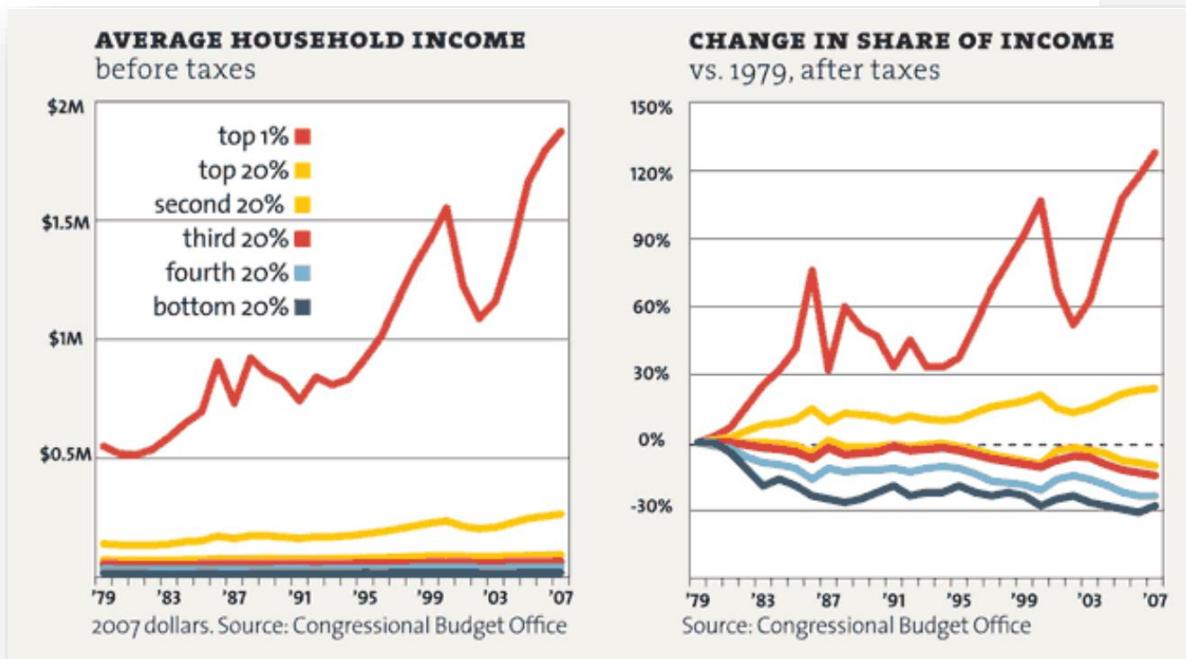
calculation of the average? The average of the *four wealthy households* is \$7,881,612; thus the average that includes the tiny-income outlier is 80% of the average for the wealthy households. There is an effect, but because of the way the arithmetic mean is computed, very small outliers have less effect on the average than very large outliers.

INSTANT TEST P 6-6

Create a list of incomes yourself in Excel. Play around with outliers to see for yourself what kind of effects they can have on the arithmetic average. Try outliers on the small side and on the large size.

Sometimes wildly deviant outliers are the result of errors in recording data or errors in transcription. However, sometimes wildly different outliers are actually rooted in reality: income inequality is observed to greater or lesser extents all over the world. For example, Figure 6-12 shows the average US pre-tax income received by the top 1% of the US population vs the quintiles (blocks of 20%) between 1979 and 2007 along with a graph showing changes in the proportion (share) of total US incomes in that period.⁶⁴ These are situations in which a geometrical mean or simply a frequency distribution may be better communicators of central tendency than a simple arithmetic mean.

Figure 6-12. Income inequality in the USA.



As with most (not all) statistics, there is a different symbol for the arithmetic mean in a sample compared with the mean of a population.

- The population mean (*parametric* mean) is generally symbolized by the Greek letter lower-case mu: μ .
- The sample mean is often indicated by a bar on top of whatever letter is being used to indicate the particular variable; thus the mean of Y is often indicated as \bar{Y} .⁶⁵

⁶⁴ (Gilson and Perot 2011)

⁶⁵ To learn how to use shortcuts to insert mathematical symbols in Word 2003, Word 2007 and Word 2010, see (Bost 2003). With these functions enabled, creating \bar{X} is accomplished by typing X followed by \bar. A curious bug is that not all fonts seem to accept the special element; for example, the Times Roman font works well, but the Garamond font (in which this textbook is mostly set) does not – the bar ends up displaced sideways over the letter. The workaround is simply to force the symbol back into a compliant font as an individual element of the text.

6.7 Median

One way to reduce the effect of outliers is to use the middle of a sorted list: the *median*. When there is an odd number of observations, there is one value in the middle where an equal number of values occur before and after that value. For the sorted list

24, 26, 26, 29, 33, 36, 37

the median is 29 because there are seven values; $7/2=3.5$ and so the fourth observation is the middle (three smaller and three larger). The median is thus 29 in this example.

When there is an even number of observations, there are two values in the middle of the list, so we average them to compute the median. For the list

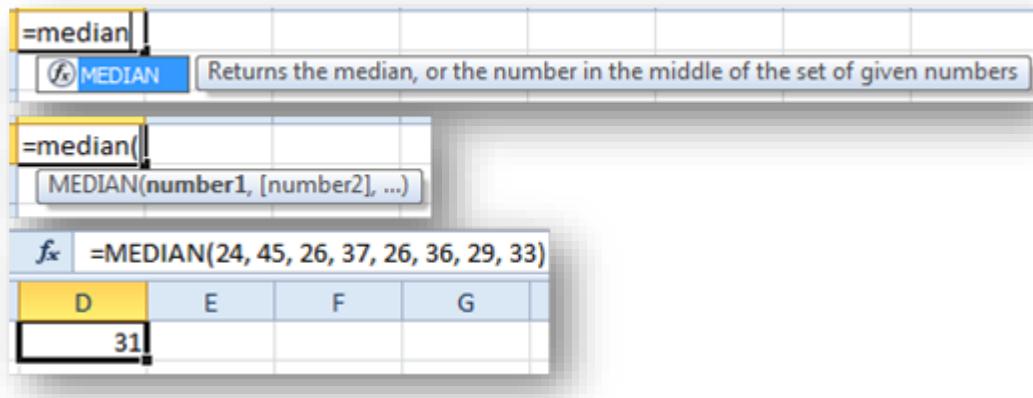
24, 26, 26, 29, 33, 36, 37, 45

the sequence number of the first middle value is $8/2 = 4$ and so the two middle values are in the fourth and fifth positions: 29 and 33. The median is $(29 + 33)/2 = 62/2 = 31$.

Computing the median by sorting and counting can become prohibitively time-consuming for large datasets; today, such computations are always carried out using computer programs such as EXCEL.

In EXCEL, the `=MEDIAN(data)` function computes the median for the data included in the cells defined by the parameter `data`, which can be individual cell addresses separated by commas or a range of cells, as shown in Figure 6-13.

Figure 6-13. MEDIAN function in Excel.



INSTANT TEST P 6-8

Create a list of values in Excel. Play around with outliers to see for yourself what kind of effects they can have on the median. Compare with the effects on the average, which you can also compute automatically. Try outliers on the small side and on the large size and note the difference in behavior of the mean and the median.

6.8 Quantiles

Several measures are related to sorted lists of data. The median is an example of a *quantile*; just as the median divides a sorted sequence into two equal parts, these *quantiles* divide the distribution into four, five, ten or 100 parts:

- *Quartiles* divide the range into four parts; the 1st quartile includes first 25% of the distribution; the second quartile is the same as the median, defining the midpoint; the third quartile demarcates 75% of the distribution below it and 25% above; the fourth quartile is the maximum of the range.
- *Quintiles* are similar to quartiles but involve five divisions.
- *Deciles* demarcate the 1st tenth of the values, the 2nd tenth and so on; the median is the 5th decile.
- *Percentiles* (sometimes called centiles) demarcate each percentage of the distribution; thus the median is the 50th percentile, the 1st quartile is the 25th percentile, the 3rd quintile is the 60th percentile and the 4th decile is the 80th percentile.

EXCEL 2010 offers several functions to locate the boundaries of the portions of a distribution in ascending rank order; however, there are subtleties in the definitions and algorithms used. Figure 6-14 lists the =PERCENTILE, =PERCENTRANK, and =QUARTILE functions. By default, we should use the .EXC versions, which are viewed as better estimators than the older versions of the functions, as discussed in §6.9.

Figure 6-14. Quantile functions in Excel 2010.

FUNCTION	DESCRIPTION
PERCENTILE.EXC function	Returns the k-th percentile of values in a range, where k is in the range 0..1, exclusive
PERCENTILE.INC function	Returns the k-th percentile of values in a range
PERCENTRANK.EXC function	Returns the rank of a value in a data set as a percentage (0..1, exclusive) of the data set
PERCENTRANK.INC function	Returns the percentage rank of a value in a data set
QUARTILE.EXC function	Returns the quartile of the data set, based on percentile values from 0..1, exclusive
QUARTILE.INC function	Returns the quartile of a data set

6.9 EXCEL 2010 .INC and .EXC Functions

In all the EXCEL 2010 quantile functions, the suffix .INC stands for *inclusive* and .EXC stands for *exclusive*.

The calculations for .INC functions are the same as for the quantile functions in EXCEL 2007 for those that exist. Functions with .INC calculations weight the positions of the estimated quantiles closer towards the median than those with the .EXC suffix, as shown in Figure 6-15 using quartiles.

The larger the sample, the smaller the difference between the two forms of computations. In general, the .EXC versions are preferred.

Figure 6-15. Comparing .INC and .EXC quantiles in Excel 2010.

Value	Seq #	Notes
121	1	min = 0th quartile.inc = 0th quartile
122	2	
123	3	
...		
144	24	
145	25	1st quartile.exc = 145.25
146	26	1st quartile.inc = 145.75 = 1st quartile
147	27	
...		
169	49	
170	50	Median = 170.5 = 2nd quartile (all)
171	51	
...		
194	74	
195	75	195.25 = 3rd quartile.inc = 3rd quartile
196	76	195.75 = 3rd quartile.exc
197	77	
...		
219	99	
220	100	max = 4th quartile.inc = 4th quartile

6.10 Quartiles in EXCEL

There are three quartile functions in EXCEL 2010, one of which (QUARTILE) matches the EXCEL QUARTILE.INC function (Figure 6-17).

The EXCEL 2010 QUARTILE and QUARTILE.INC functions are identical to the EXCEL 2007, 2013 and 2016 QUARTILE function. Figure 6-18 shows the comparison of results of these types of quartiles for a sample dataset.

Figure 6-17. Excel 2010 quartile functions.

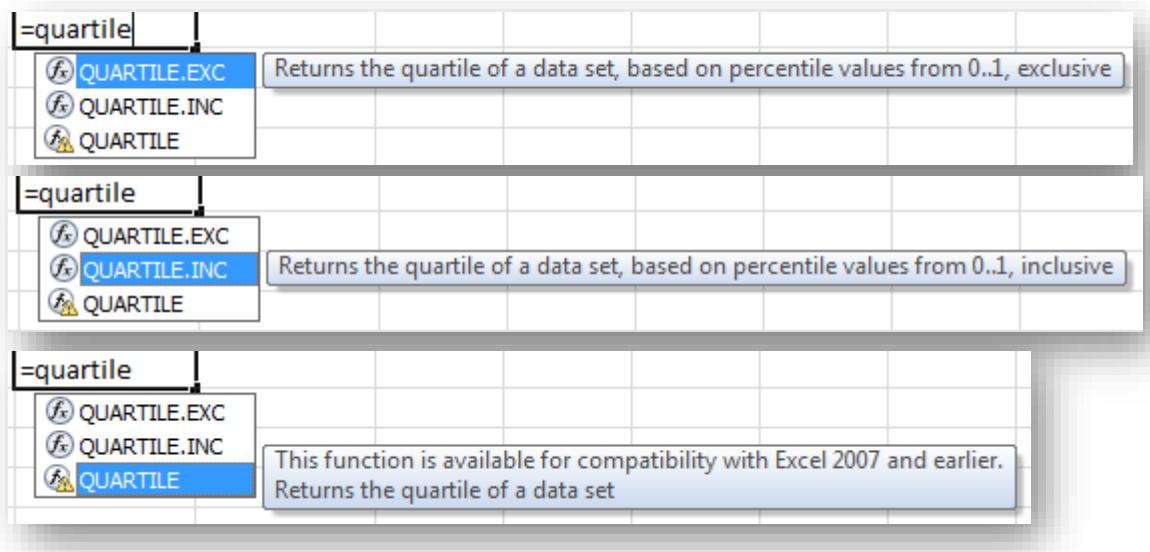


Figure 6-18. Results of Excel 2010 quartile functions.

	E	F	G	H
1	QUARTILE FUNCTIONS			
2	PARM	QUARTILE.EXC	QUARTILE.INC	QUARTILE (2003-7)
3	0	#NUM!	121	121
4	1	145.25	145.75	145.75
5	2	170.5	170.5	170.5
6	3	195.75	195.25	195.25
7	4	#NUM!	220	220

Figure 6-16. Formulas for comparison of Excel 2010 quartile functions.

	E	F	G	H
1	QUARTILE FUNCTIONS			
2	PARM	QUARTILE.EXC	QUARTILE.INC	QUARTILE (2003-7)
3	0	=QUARTILE.EXC(\$A\$2:\$A\$101,\$E3)	=QUARTILE.INC(\$A\$2:\$A\$101,\$E3)	=QUARTILE(\$A\$2:\$A\$101,\$E3)
4	1	=QUARTILE.EXC(\$A\$2:\$A\$101,\$E4)	=QUARTILE.INC(\$A\$2:\$A\$101,\$E4)	=QUARTILE(\$A\$2:\$A\$101,\$E4)
5	2	=QUARTILE.EXC(\$A\$2:\$A\$101,\$E5)	=QUARTILE.INC(\$A\$2:\$A\$101,\$E5)	=QUARTILE(\$A\$2:\$A\$101,\$E5)
6	3	=QUARTILE.EXC(\$A\$2:\$A\$101,\$E6)	=QUARTILE.INC(\$A\$2:\$A\$101,\$E6)	=QUARTILE(\$A\$2:\$A\$101,\$E6)
7	4	=QUARTILE.EXC(\$A\$2:\$A\$101,\$E7)	=QUARTILE.INC(\$A\$2:\$A\$101,\$E7)	=QUARTILE(\$A\$2:\$A\$101,\$E7)

6.11 QUARTILE.EXC vs QUARTILE.INC

The following information is provided for students who are curious about the different calculation methods used for n-tiles in EXCEL. The details are unnecessary if the advice "use the .EXC versions" is acceptable at face value.

Depending on the number of data points in a sample, it may be necessary to estimate ("interpolate") between existing values when computing quartiles. For example, if we have exactly 10 values in the sample, we have to compute the second quartile (Q2, which is the median) as being half way between the observed values #5 and #6 in the ranked list if we label the minimum as #1. If we label the minimum as #0, Q2 is half way between the values called #4 and #5. The calculations of Q1 and Q3 depend on whether we start counting the minimum as #0 or as #1.

Newer versions of EXCEL include a function called =QUARTILE.EXC where EXC stands for exclusive. In this method, the minimum is ranked as #1 and the maximum in our sample of 10 sorted values is called #10. The median (Q2) is half way between value #5 and value #6. The value of Q1 is computed as if it were the $0.25*(N+1)$ th value (remembering that the minimum is labeled #1); in our example, that would be the observation ranked as #2.75 – that is, $3/4$ of the distance between observation #2 and observation #3.. Similarly, Q3 is computed as the $0.75*(N+1)$ th = $0.75*11$ which would be rank #8.25, $1/4$ of the way between observation #8 and observation #9. There are 2 values below Q1 in our sample of 10 values and 2 values above Q3. There are 3 values between Q1 and Q2 and 2 value between Q2 and Q3.

The older versions of EXCEL use the =QUARTILE function which is exactly the same as the modern =QUARTILE.INC function. The INC stands for inclusive and the minimum is labeled as #0. This method is the commonly used version of the calculations for quartiles. In this method, for the list of 10 values (#0 through #9), the median (Q2) is halfway between the values ranked as #4 and #5. Q2 is calculated as if it were rank $\#(0.25*(N-1)) = \#2.25$ – that is, 25% of the distance between value ranked #2 and the one that is #3. Similarly, Q3 is the value corresponding to rank $\#(0.75*(N-1)) = \#6.75$ – 75% of the distance between values #6 and #7. Thus in our example there are 3 values below Q1 in our sample of 10 values and 3 values above Q3. There are 2 values between Q1 and Q2 and 2 values between Q2 and Q3.

The QUARTILE.EXC function is theoretically a better representation of the values of Q2 and of Q3. You can simply ignore the QUARTILE.INC and QUARTILE functions in most applications, but they are available if you are ordered to use it.⁶⁶

⁶⁶ For detailed explanation of the differences between these functions, see (Alexander 2013) <<http://datapigtechnologies.com/blog/index.php/why-excel-has-multiple-quartile-functions-and-how-to-replicate-the-quartiles-from-r-and-other-statistical-packages/>> or <<http://tinyurl.com/kcrxlfn>>.

6.12 Box Plots

A common graphical representation of data distributions is the *box plot*, also called a *box-and-whisker diagram*, shown for three datasets (Figure 6-19) in the diagram (Figure 6-20).

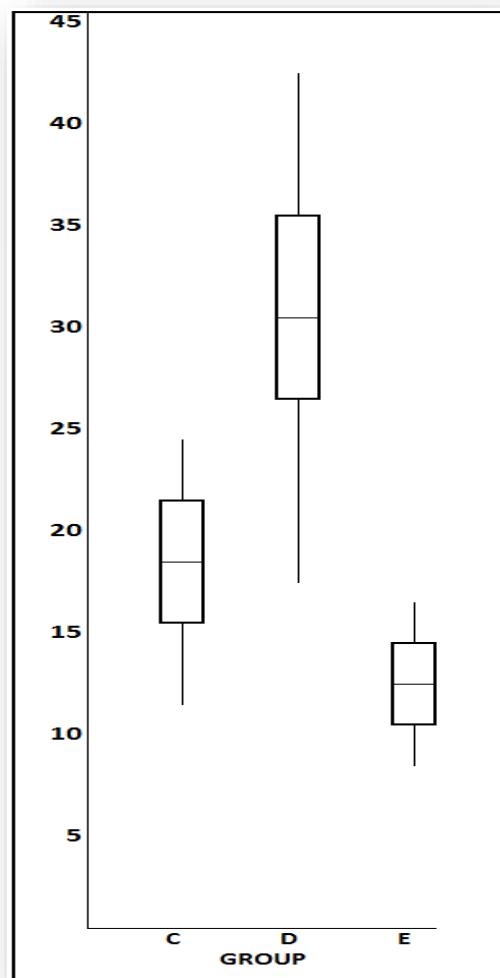
Figure 6-19. Sample data for box-and-whisker plots.

Statistic	Concentran	Denmartox	Ephrumia
max	24	42	16
q3	21	35	14
median	18	30	12
q1	15	26	10
min	12	17	8

As you can see, the top “whisker” (vertical line) runs from the maximum to the third quartile; the bottom whisker runs from the first quartile to the minimum. The box runs from the third quartile through the median (horizontal line inside the box) down to the first quartile.

Unfortunately, no version of EXCEL yet provides an automatic method for drawing box-and-whisker plots. However, if one has a small number of categories, highlighting the individual cells to draw borders (for lines) and boxes is not difficult.⁶⁷

Figure 6-20. Box-and-whisker plots for sample data.



⁶⁷ (Peltier 2011)

6.13 Percentiles in EXCEL

The same principles apply to percentiles as to quartiles.

Figure 6-14 lists the =PERCENTILE.EXC and the =PERCENTILE.INC functions. Both produce an estimate of the value in an input array that corresponds to any given percentile. The .EXC functions are recommended.

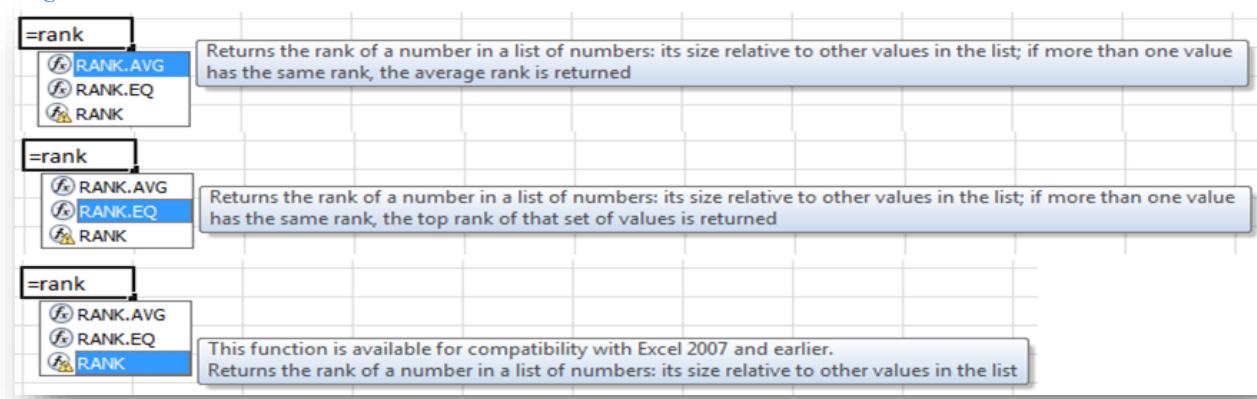
The HELP facility provides ample information to understand and apply the =PERCENTILE functions.

6.14 Rank Functions in EXCEL

You can produce a ranked list (ascending or descending) easily in EXCEL 2010. Using a =RANK, =RANK.AVG, or =RANK.EQ function (Figure 6-21), you can compute a rank for a given datum in relation to its dataset.

For example, Figure 6-22 shows information from the quality-control section of GalaxyFleet for its 17 classes

Figure 6-21. Rank functions in Excel 2010.



of starships (Andromeda Class, Betelgeuse Class, etc.). =RANK.EQ is the EXCEL 2010 version of the older EXCEL versions' =RANK function. The data don't have to be sorted to be able to compute ranks. In this case, the data are simply ordered alphabetically by class.

INSTANT TEST P 6-14

Create some data using one of the random-number generator functions { =RAND(), =RANDBETWEEN(bottom, top)}.
Use the =RANK.EQ and =RANK.AVG functions and compare the results.

Explain these results as if to someone who has never heard of the functions.

Figure 6-22. Accident data sorted by starship class.

	Total Number of Accidents per Month in GalaxyFleet Starships Classes	RANK (older version)	RANK.EQ	RANK.AVG
Andromeda	142	16	16	16
Betelgeuse	123	10	10	10
Chara	107	4	4	4
Deneb	126	11	11	11
Eltanin	97	2	2	2
Fomalhaut	134	12	12	12
Gomeisa	122	8	8	8.5
Hamal	121	6	6	6.5
Jabbah	157	17	17	17
Kornephoros	140	13	13	14
Lesuth	121	6	6	6.5
Marfik	122	8	8	8.5
Nihal	100	3	3	3
Pollux	140	13	13	14
Rigel	94	1	1	1
Sargas	140	13	13	14
Thuban	120	5	5	5

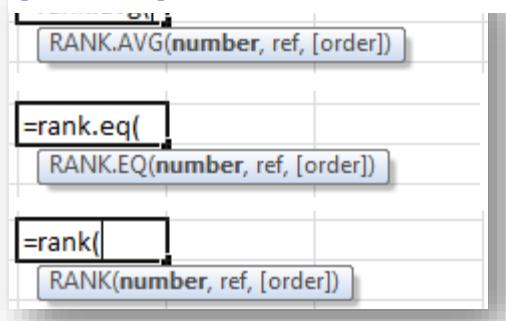
In case of ties, the =RANK and =RANK.EQ functions both choose the higher rank whereas the =RANK.AVG function computes the average of the tied ranks. Figure 6-23 shows the results of the three rank functions starting rank #1 at the lowest value. Notice that Hamal Class and Lesuth Class starships had the same value and therefore could be ranks 6 and 7. The =RANK.EQ and =RANK functions list them as rank 6 (and thus rank 7 is not listed) whereas the =RANK.AVG function averages the ranks [$(6+7)/2 = 6.5$] and shows that value (6.5) for both entries.

Figure 6-23. Rank functions in Excel 2010.

Total Number of Accidents per Month in GalaxyFleet Starships Classes	RANK (older version)	RANK.EQ	EQ v AVG	RANK.AVG
Rigel	94	1	1	= 1
Eltanin	97	2	2	= 2
Nihal	100	3	3	= 3
Chara	107	4	4	= 4
Thuban	120	5	5	= 5
Hamal	121	6	6	< 6.5
Lesuth	121	6	6	< 6.5
Gomeisa	122	8	8	< 8.5
Marfik	122	8	8	< 8.5
Betelgeuse	123	10	10	= 10
Deneb	126	11	11	= 11
Fomalhaut	134	12	12	= 12
Kornephoros	140	13	13	< 14
Pollux	140	13	13	< 14
Sargas	140	13	13	< 14
Andromeda	142	16	16	= 16
Jabbah	157	17	17	= 17

Each rank function in EXCEL 2010 requires the specific datum to be evaluated (**number**), the reference dataset (**ref**) and an optional **order** parameter, as shown in Figure 6-24.

Figure 6-24. Excel 2020 rank functions showing optional order parameter.



By default, the **order** parameter (Figure 6-26) is zero and therefore, if it is not entered, the value corresponding to rank #1 is the maximum value in the reference list. Figure 6-25 shows the results of a default-order sort and the formulas used to generate the ranks in EXCEL 2010.

Figure 6-26. Optional *order* parameter in Excel 2010 rank functions.

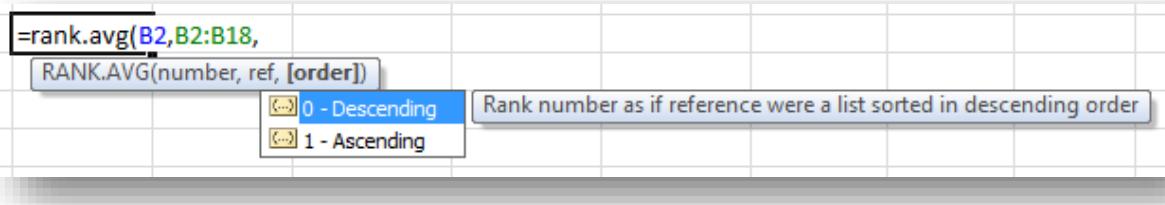


Figure 6-25. Rank functions (values in top table and formulas in bottom table) using default sort order.

Total Number of Accidents per Month in GalaxyFleet Starships Classes	RANK (older version)	RANK.EQ	EQ v AVG	RANK.AVG
Rigel	94	17	=	17
Eltanin	97	16	=	16
Nihal	100	15	=	15
Chara	107	14	=	14
Thuban	120	13	=	13
Hamal	121	11	<	11.5
Lesuth	121	11	<	11.5
Gomeisa	122	9	<	9.5
Marfik	122	9	<	9.5
Betelgeuse	123	8	=	8
Deneb	126	7	=	7
Fomalhaut	134	6	=	6
Kornephoros	140	3	<	4
Pollux	140	3	<	4
Sargas	140	3	<	4
Andromeda	142	2	=	2
Jabbah	157	1	=	1

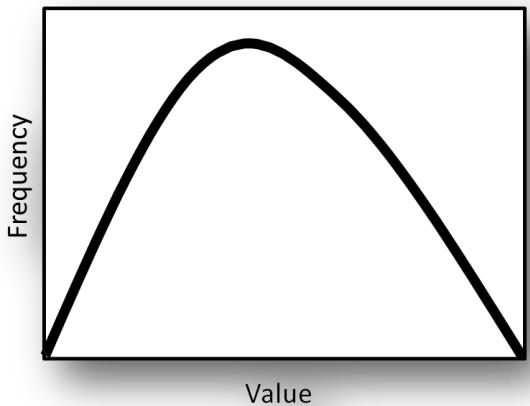
Total Number of Accidents per Month in GalaxyFleet Starships Classes	RANK()	RANK.EQ	RANK.AVG
Rigel	=RANK(\$B2,\$B\$2:\$B\$18)	=RANK.EQ(\$B2,\$B\$2:\$B\$18)	=RANK.AVG(\$B2,\$B\$2:\$B\$18)
Eltanin	=RANK(\$B3,\$B\$2:\$B\$18)	=RANK.EQ(\$B3,\$B\$2:\$B\$18)	=RANK.AVG(\$B3,\$B\$2:\$B\$18)
Nihal	=RANK(\$B4,\$B\$2:\$B\$18)	=RANK.EQ(\$B4,\$B\$2:\$B\$18)	=RANK.AVG(\$B4,\$B\$2:\$B\$18)
Chara	=RANK(\$B5,\$B\$2:\$B\$18)	=RANK.EQ(\$B5,\$B\$2:\$B\$18)	=RANK.AVG(\$B5,\$B\$2:\$B\$18)
Thuban	=RANK(\$B6,\$B\$2:\$B\$18)	=RANK.EQ(\$B6,\$B\$2:\$B\$18)	=RANK.AVG(\$B6,\$B\$2:\$B\$18)
Hamal	=RANK(\$B7,\$B\$2:\$B\$18)	=RANK.EQ(\$B7,\$B\$2:\$B\$18)	=RANK.AVG(\$B7,\$B\$2:\$B\$18)
Lesuth	=RANK(\$B8,\$B\$2:\$B\$18)	=RANK.EQ(\$B8,\$B\$2:\$B\$18)	=RANK.AVG(\$B8,\$B\$2:\$B\$18)
Gomeisa	=RANK(\$B9,\$B\$2:\$B\$18)	=RANK.EQ(\$B9,\$B\$2:\$B\$18)	=RANK.AVG(\$B9,\$B\$2:\$B\$18)
Marfik	=RANK(\$B10,\$B\$2:\$B\$18)	=RANK.EQ(\$B10,\$B\$2:\$B\$18)	=RANK.AVG(\$B10,\$B\$2:\$B\$18)
Betelgeuse	=RANK(\$B11,\$B\$2:\$B\$18)	=RANK.EQ(\$B11,\$B\$2:\$B\$18)	=RANK.AVG(\$B11,\$B\$2:\$B\$18)
Deneb	=RANK(\$B12,\$B\$2:\$B\$18)	=RANK.EQ(\$B12,\$B\$2:\$B\$18)	=RANK.AVG(\$B12,\$B\$2:\$B\$18)
Fomalhaut	=RANK(\$B13,\$B\$2:\$B\$18)	=RANK.EQ(\$B13,\$B\$2:\$B\$18)	=RANK.AVG(\$B13,\$B\$2:\$B\$18)
Kornephoros	=RANK(\$B14,\$B\$2:\$B\$18)	=RANK.EQ(\$B14,\$B\$2:\$B\$18)	=RANK.AVG(\$B14,\$B\$2:\$B\$18)
Pollux	=RANK(\$B15,\$B\$2:\$B\$18)	=RANK.EQ(\$B15,\$B\$2:\$B\$18)	=RANK.AVG(\$B15,\$B\$2:\$B\$18)
Sargas	=RANK(\$B16,\$B\$2:\$B\$18)	=RANK.EQ(\$B16,\$B\$2:\$B\$18)	=RANK.AVG(\$B16,\$B\$2:\$B\$18)
Andromeda	=RANK(\$B17,\$B\$2:\$B\$18)	=RANK.EQ(\$B17,\$B\$2:\$B\$18)	=RANK.AVG(\$B17,\$B\$2:\$B\$18)
Jabbah	=RANK(\$B18,\$B\$2:\$B\$18)	=RANK.EQ(\$B18,\$B\$2:\$B\$18)	=RANK.AVG(\$B18,\$B\$2:\$B\$18)

6.15 Mode(s)

Frequency distributions often show a single peak, called the *mode*, that corresponds in some sense to the most popular or otherwise most frequent class of observations. However, it is also possible to have more than one mode.

Figure 6-27 shows a typical unimodal frequency distribution – one with a single class that is the most frequent.

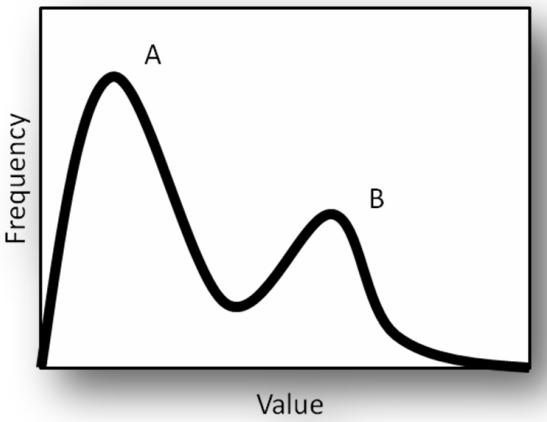
Figure 6-27. Unimodal frequency distribution.



Finding such a class by inspection is easy: just look at the table of frequencies or examine the graph of the frequency distribution.

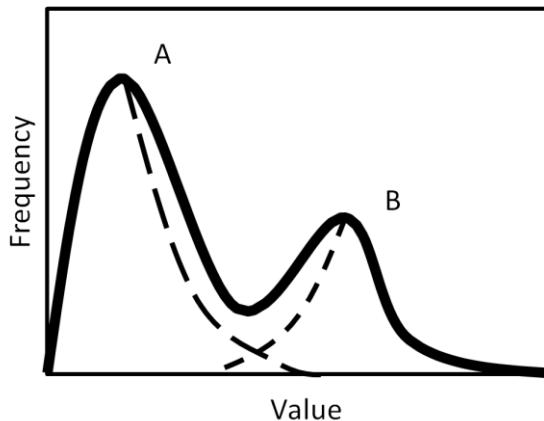
Figure 6-28 shows a more difficult problem: a distribution with more than one *local maximum*. The peak labeled A is definitely a mode but peak B can justifiably be called a secondary mode as well for this bimodal distribution. B marks a region that is more frequent than several adjacent regions, so it may be important in understanding the phenomenon under study.

Figure 6-28. Bimodal frequency distribution.



Sometimes a bimodal distribution is actually the combination of two different distributions, as shown by the dashed lines in Figure 6-29. For example, perhaps a financial analyst has been combining performance data for stocks without realizing that they represent two radically different industries, resulting in a frequency distribution that mixes the data into a bimodal distribution.

Figure 6-29. Bimodal distribution resulting from combination of two underlying distributions.



Generalizing about data that mix different underlying populations can cause errors, misleading users into accepting grossly distorted descriptions, and obscuring the phenomena leading to the differences. One of the situations in which such errors are common are political discussions of income inequality, where statements about overall average changes in income are largely meaningless. Only when specific demographic groups which have very different patterns of change are examined can the discussion be statistically sound. Similarly, some diseases have radically different properties in children and adults or in men and women; ignoring these differences may generate odd-looking frequency distributions that do not fit the assumptions behind general descriptions of central tendency and of dispersion.

As you continue your study of statistics in this and later courses, you will see that many techniques have been developed precisely to test samples for the possibility that there are useful, consistent reasons for studying different groups separately. Examples include multiway analysis of variance (ANOVA) to test for systematic influences from different classifications of the data (men, women; youngsters, adults; people with different blood types; companies with different management structures; buildings with different architecture) on the observations.

INSTANT TEST P 6-18

Create a column of 100 rows of data using function =INT(NORM.INV(RAND(),100,10)) and then add the same number of entries using =INT(NORM.INV(RAND(),150,10)).

Construct a frequency distribution of the combined data and explain the results.

6.16 Statistics of Dispersion

Observations may have the same arithmetic mean yet obviously be different in how widely the data vary.

Figure 6-30 shows three frequency distributions with the same mean and sample sizes but different distribution (*dispersion*) patterns.

The common ways of describing the extent of dispersion are the *range*, the *variance*, the *standard deviation*, and the *interquartile range*.

6.17 Range

As you've seen in several discussions before this, the range is simply the difference between the maximum value and the minimum value in a data set. Thus if a rank-ordered data set consists of $\{3, 4, 4, 8, \dots, 22, 24\}$ ⁶⁸ then its range is $24 - 3 = 21$.

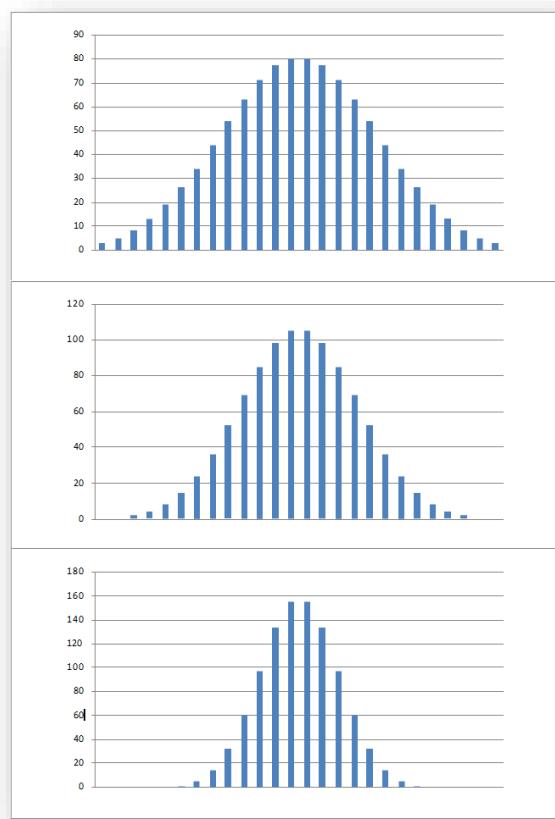
As discussed in §6.1⁶⁹ (Summarizing Groups of Data using EXCEL Descriptive Statistics), the **Data | Data Analysis | Descriptive Statistics** tool generates a list of descriptive statistics that includes the range. EXCEL has no explicit function for the range, but it's easy simply to compute the maximum minus the minimum, as shown in Figure 6-31.

Figure 6-31. Computing the range in Excel.

f_x	=MAX(A1:A15)-MIN(A1:A15)
D	E
93	

The descriptive statistics discussed in section 6.1 automatically calculate the range.

Figure 6-30. Three different frequency distributions with same mean but different dispersion.



INSTANT TEST P 6-19

Using data similar to those you created in the test on the previous page, calculate the range of your data.

⁶⁸ By convention a set (group) is enclosed in braces {} in text; commas separate individual elements and colons indicate a range (e.g., 4:8). Similar conventions apply to Excel, except that arguments of functions are in parentheses () .

⁶⁹ The symbol § means "section" and is used for references to numbered section headings in this textbook.

6.18 Variance: σ^2 and s^2

The variance, σ^2 (*sigma squared*), is used throughout applied statistics. It is the *average of squared deviations from the mean*.

We start by computing the individual deviations y , from the mean:

$$y = Y - \bar{Y}$$

where y is a *deviate*.

In the early years of applied statistics, statisticians tried using the average of these deviates as a measure of dispersion. However, the sum of all the deviates around the mean is always zero, so they tried using the absolute value of y , represented as $|y|$. That worked, but the average deviate turned out not to be a reliable measure of dispersion. The statisticians developed the idea of *squaring* the deviates before averaging them, resulting in the *variance*, represented for an entire group (what we call the *parametric variance*) as σ^2 . Thus the equation for the parametric variance is as shown below:

$$\sigma^2 = \frac{\sum y^2}{n}$$

However, calculating individual deviates, y , is tedious, so a better formula for parametric variance is derived from expanding the definition of y :

$$\Sigma y = \Sigma (Y - \bar{Y})^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} \quad \text{and so} \quad \sigma^2 = [\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}] / n$$

If our data set consists only of randomly selected values from all possible observations – what we call a *sample* – then we have to use a slightly different formula for the sample variance. The reason is that when we compute a statistic based on a sample, we expect to that statistic to have the same value on average as the parametric statistic (the one for the entire population). We say that a good sample statistic is an *unbiased estimator* of the parametric value of that statistic.

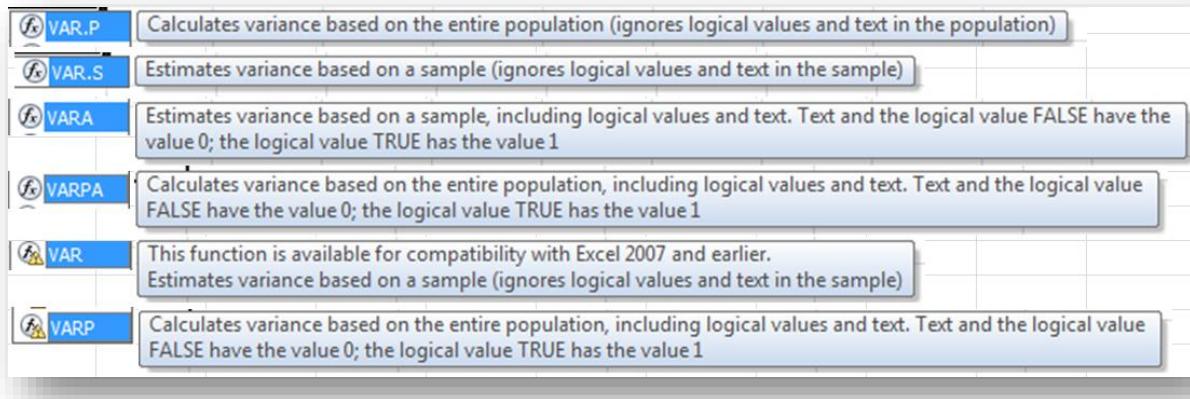
It turns out that the parametric variance calculation is slightly too small: variances calculated on samples consistently underestimate the parametric variance. The correct, unbiased estimator of the variance is the sample variance, s^2 ; it is calculated almost the same way as σ^2 except that we divide the sum of the squared deviates by one less than the sample size, thus producing a slightly larger value than that of σ^2 on the same data:

$$s^2 = \frac{\sum y^2}{n-1}$$

This is the statistic that is most commonly used to describe the dispersion or variability of a data set, since most data sets are samples.

Practically no one actually uses manual calculation of variance today, however. EXCEL, for example, has several variance functions, as shown in Figure 6-32, which is a composite image that shows all the pop-up descriptors at once (normally one sees only one at a time). With your experience of other EXCEL functions, you can now easily learn about these using the EXCEL help functions.

Figure 6-32. Excel 2010 variance functions.



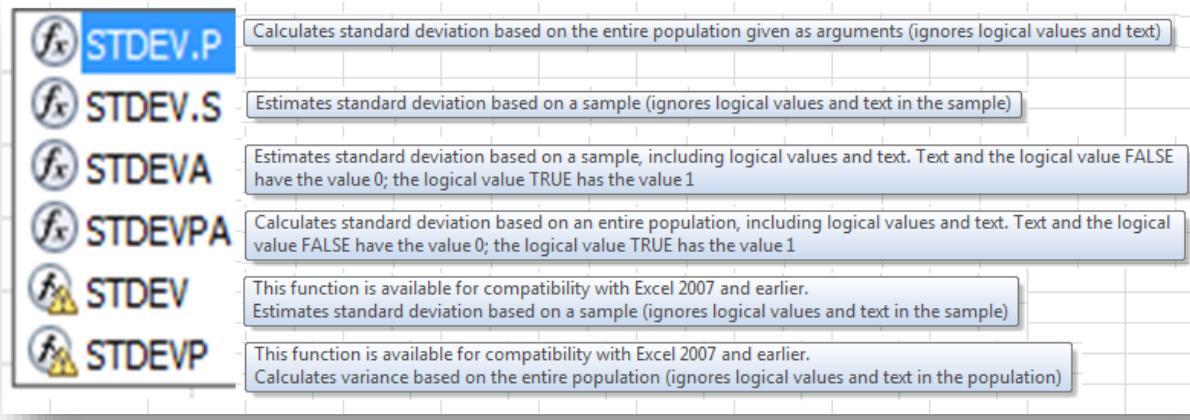
6.19 Standard Deviation: σ and s

The standard deviation is simply the square root of the variance:

$$\sigma = \sqrt{\sigma^2} \text{ and } s = \sqrt{s^2}$$

EXCEL 2010 functions are shown in a composite image in Figure 6-33.

Figure 6-33. Excel 2010 standard-deviation functions.



The standard deviation is used extensively in computations of *confidence intervals* and *confidence limits* in *statistical estimation*. It also plays an important role in many *hypothesis tests* about whether observed sample statistics support theoretical models about the corresponding parametric statistics (e.g., testing to see if samples support the hypothesis of equality of parametric means among groups).

6.20 Skewness

The output from the **Descriptive Statistics** tool in **Data Analysis** includes coefficients of *kurtosis* and of *skewness*, as shown in Figure 6-4, which is reproduced in Figure 6-34 with dots to highlight the relevant entries.

Skewness is a measure of asymmetry:

- A *negative* skewness coefficient (*skewed to the left*) indicates that more than half of the data lie to the *left* of the mean (Figure 6-35)⁷⁰ – sometimes (but not always) pulling the median to the left of the mean.

Figure 6-35. Frequency distribution showing negative skewness coefficient.

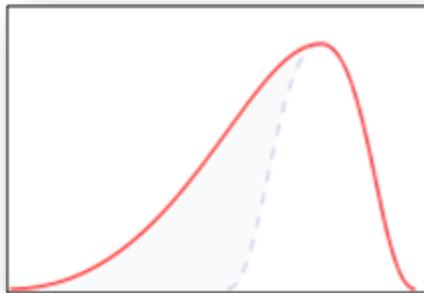


Figure 6-34. Descriptive statistics highlighting kurtosis and skewness coefficients.

Network Attacks	
Mean	2844.425
Standard Error	25.64544
Median	2859
Mode	2872
Standard Deviation	489.9554
Sample Variance	240056.3
Kurtosis	-0.14605
Skewness	-0.03163
Range	2552
Minimum	1632
Maximum	4184
Sum	1038215
Count	365
Largest(5)	3936
Smallest(5)	1692
Confidence Level(95.0%)	50.43182

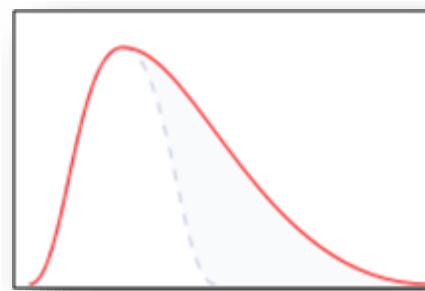
- A *zero* skewness coefficient indicates a symmetrical curve, with equal numbers of observations to the left and the right of the mean, allowing the median and the mean to be similar or identical;
- A *positive* skewness coefficient (*skewed to the right*) indicates that more than half of the data lie to the *right* of the mean (Figure 6-36) – sometimes (but not always) pulling the median to the right of the mean.

The EXCEL function =SKEW(range) generates the *sample* skewness coefficient, g_1 , which estimates a parametric skewness coefficient denoted γ_1 . It is possible to modify the result to compute a parametric γ_1 but that level of detail is unnecessary in this introductory course.

INSTANT TEST P 6-22

Using generated data, practice using all the functions discussed in this section, including the kurtosis and skew coefficients on the next pages.

Figure 6-36. Frequency distribution showing positive skewness coefficient.



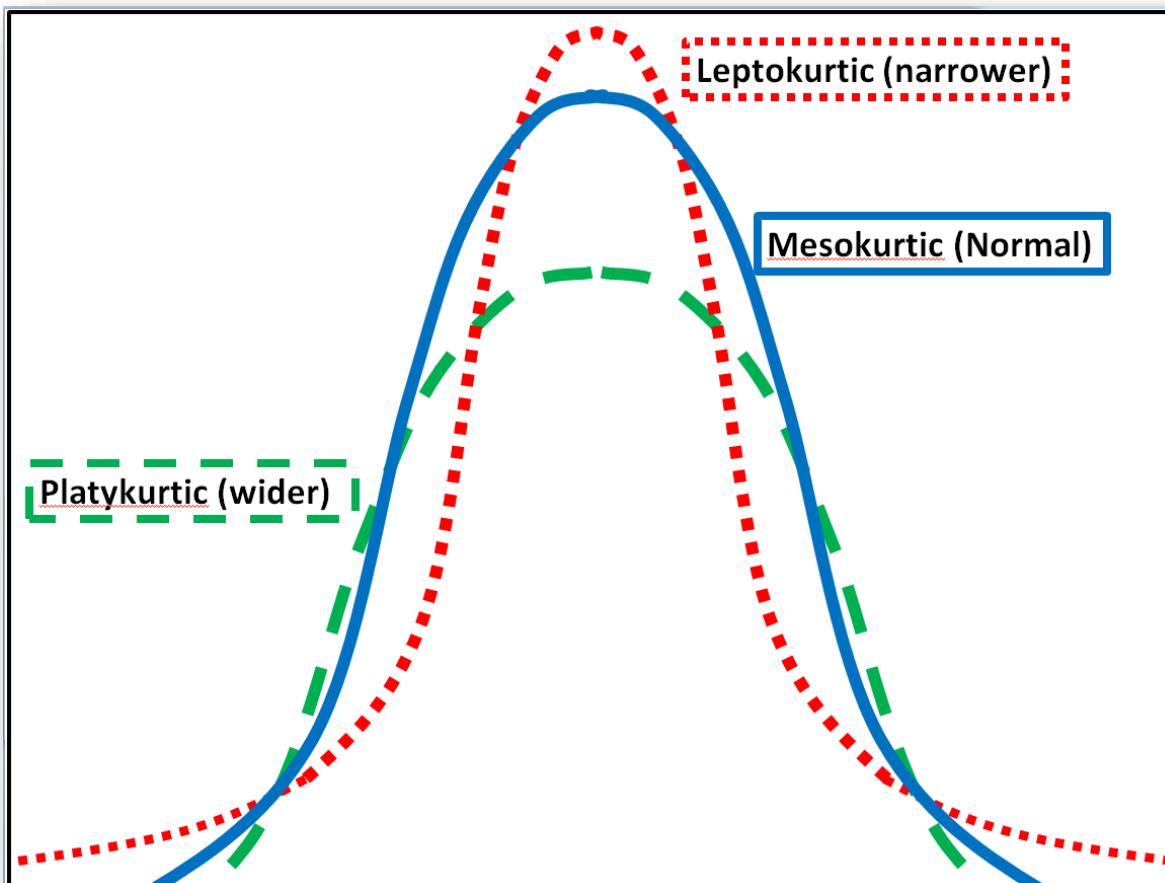
⁷⁰ The two illustrations of skewness are based on the *Wikimedia Commons* file < Skewness Statistics.svg > freely available for use with attribution from <http://en.wikipedia.org/wiki/File:Skewness_Statistics.svg>.

6.21 Kurtosis

The coefficient of *kurtosis* (g_2 for samples and γ_2 for populations) describes how wide or narrow (sometimes described as how *peaked*) an observed distribution is compared to the Normal distribution.

- A *platykurtic* distribution⁷¹ (g_2 or $\gamma_2 < 0$) is shorter than the Normal distribution in the middle, has more observations in the two shoulders, and has fewer observations in the tails;
- A *mesokurtic* distribution⁷² (g_2 or $\gamma_2 = 0$) is exemplified by the Normal distribution itself;
- A *leptokurtic* distribution⁷³ (g_2 or $\gamma_2 > 0$) is taller than the Normal distribution in the middle, has fewer observations in the two shoulders, and has more observations in the tails.

Figure 6-37. Leptokurtic, mesokurtic and platykurtic frequency distributions.



In EXCEL, the =KURT(range) function returns the sample coefficient of kurtosis, g_2 , which estimates the parametric value γ_2 for an array. Just as in the discussion of the skewness coefficient, it is possible to modify the result to compute a parametric γ_2 but that level of detail is unnecessary in this introductory course.

⁷¹ From Greek πλατύς ζ = platus = flat & κυρτός ζ = kurtosis = curvature

⁷² From Greek μέσος ζ = mesos = middle

⁷³ From Greek λεπτός ζ = leptos = small

7 Sampling and Statistical Inference

7.1 Populations and Samples

In several sections of your introduction so far, you have read about *parametric* statistics and *sample* statistics. In this section we examine these concepts – populations and samples – in more depth.

When the data in which we are interested represent everything on which we are focusing, we call the data set a *population*. For example, we could discuss how the students in the QM213 Business & Economic Statistics I course in the School of Business and Management in the Fall semester of 2029 at Norwich University do in their first quiz and in their final exam. These data could constitute the entire population. If we consider the data the population, then the mean, standard deviation, median, mode and any other statistic we compute using these data are all *parametric* values. They are not *estimates* of anything; they are exact representations of the population attributes for that particular, specific group. The group doesn't represent anything; it isn't intended to be anything other than itself.

Similarly, if we summarize the sports statistics for a particular team for a specific period, the data are not a sample of anything: they are the entire population of data for that period. "The Norwich Paint-Drying Team scored an average of 32 points per round in the North American Paint-Drying Tourney in 2029" isn't a description of a sample: it's (presumably) the absolute, incontrovertible truth; it's a parametric statistic.

But what if we consider the QM213 data as part of a larger study? What if we are actually interested in studying the relationship between the score on the first quiz in QM213 and the score on the final exam in QM213 classes in general? The data we just collected could actually be going into a collection spanning the years from, say 2010 through 2029; in that case, the group of interest is not only the particular class and the particular quiz results: the group of interest is *all possible groups of QM213 students* and their first quiz and final exam results. In this case, the data for the Fall 2029 QM213 class's first quiz and final exam results are both *samples* from the larger, theoretical *population* of all possible such values.

As stated above, does the population include students in QM213 courses before 2010 and after 2029? Does the population for which the Fall 2029 results have been collected include Spring QM213 classes? Does the population include other statistics classes in the School of Business and Management such as QM 370 *Quantitative Methods in Marketing & Finance*? Does the population include results from the first quiz and final exams for other statistics courses at Norwich University such as MA232 *Elementary Statistics*? Does it include results for statistics courses at other universities? For that matter, is the population we are studying all possible courses that have a first quiz and a final exam?

The critical concept here is that there is nothing absolute about a set of numbers that tells us instantly whether they are a sample or a population; there's no convenient little flag sticking up to indicate that status. More seriously, the decision on whether to view a group of data as a sample or a population is not based on the data: the decision is based on *the way the data are collected* and *how they are being used* by the analysts.

7.2 Sample Statistics and Parameters

One of the most important concepts in statistics is the idea of representative samples. A sample is representative when the information from the sample can be used to guess at the values of the population from which it was drawn. We say that we can infer the parametric value of a statistic from the value of the sample statistic.

A researcher could claim that the Fall 2010 Norwich University QM213 quiz and final scores were samples from the global population of all statistics courses given anywhere at any time. There would be a number of assumptions in such a claim. Think about some of the claims the researcher could be making by asserting that the sample in question was from the population of all students taking any statistics course (this is only a partial list):

- The QM213 class in Fall 2029 is similar to all other QM213 classes;
- The QM213 course is similar to all other statistics courses the School of Business and Management;
- The statistics courses in the School of Business and Management' to all other statistics courses at Norwich University;
- Statistics courses at Norwich University are similar to all other statistics courses on planet Earth;
- Statistics courses on planet Earth are similar to all other statistics courses in the known universe.

None of these assumptions is obligatory; what the researcher decides to claim about the nature of the population from which the Fall 2029 QM213 first quiz and final exam results determines what assumptions are being made.

Depending on what the population is assumed to be, the researcher will be able to try to infer attributes of that population based on the particular sample; whether those inferences will be accepted by other statisticians is a question of how carefully the researcher thinks about the sampling process.

In ordinary life, we are faced with data that are represented to be from populations defined according to the preferences of the people reporting the statistics. For example, a newspaper article may report that 23% of the college students in Mare Imbrium have been unable to buy a hovercraft in the first year after their graduation. The writer goes on to discuss the general problems of college graduates system wide, including those on Earth, on the Lunar Colonies, and on the Jovian Satellite Colonies. But how do we know that the Mare Imbrium students are in fact from the population defined for the entire Solar System? Is there any evidence presented to suggest that the Mare Imbrium students are in fact a representative sample? What about the fact that the proportion of college graduates who buy a hovercraft on Mars has reached 91%? Or that Earth graduates have a paltry 3% ownership of these much-desired vehicles in their first year after graduation?

Whenever you read statistics, especially in the popular press or in media prepared by people with a strong interest in convincing you of a particular point of view, you should investigate in depth just how the data were collected and on what basis they can be considered representative of the population for which they are claimed to be samples.

INSTANT TEST P 2

Examine published reports that include statistical information. Notice carefully where the authors are discussing *populations* and where they are discussing *samples*. Think about what you would have to do to extend or narrow the definitions to change the populations to larger or smaller groups. Explain your reasoning as if to a fellow student.

7.3 Greek Letters for Parametric Statistics

You will also have noticed in previous sections that parametric statistics are customarily symbolized using lowercase Greek letters. For reference, Figure 7-1 shows the Greek alphabet with names and Roman equivalents. Notice that sigma, the equivalent of our *s*, has two lowercase versions, σ and ς . The latter is rarely used in mathematics; it is the form that is used in Greek writing only for a sigma that is at the end of a word.

Figure 7-1. Greek letters for parametric statistics.

A α	B β	Γ γ	Δ δ
Alpha/a	Beta/b	Gamma/g	Delta/d
E ε	Z ζ	H η	Θ θ
Epsilon/ě	Zeta/z	Eta/ē	Theta/th
I ι	K κ	Λ λ	M μ
Iota/I	Kappa/k	Lambda/l	Mu/m
N ν	Ξ ξ	O \circ	Π π
Nu/n	Xi/x	omicron/ő	pi/p
P ρ	Σ σ ς	T τ	Y υ
Rho/r	Sigma/s	Tau/t	Upsilon/u
Φ ϕ	X χ	Ψ ψ	Ω ω
Phi/ph	chi/ch	psi/ps	omega/ő

7.4 Random Sampling from a Population

What makes a sample representative of a particular population?

Should we inspect the data and pick the ones we think look like what we believe the population to be? Bad idea, don't you think? How would we ever separate the effects of our own preconceptions from the reality of the situation? With a pick-and-choose approach to sampling, we could claim anything we wanted to and pretend to provide a statistical justification for our claims.

For example, suppose a (shall we say) naïve researcher, Arthur Schlemiel,⁷⁴ has a preconceived notion that the score on the first quiz in the QM213 class for Fall 2029 was strongly related to the score on the final exam in that class. Figure 7-2 shows the original data.

Schlemiel could cheerfully (and wrongly) select only the students whose Quiz #1 scores and Final Exam scores were similar; for example, either both low, both middling, or both high. Schlemiel might compute the ratio of the Final Exam score to the Quiz #1 score (F/Q) and pick only the students with a F/Q ratio of, say, 85% to 115%. The students whose data are in bold italics and are shown in the central box in Figure 7-3.

Figure 7-3. Biased sampling using F/Q ratio.

Student	Quiz #1	Final Exam	F/Q
P	93%	64%	69%
N	74%	60%	81%
H	98%	82%	84%
E	90%	82%	91%
J	98%	97%	99%
Q	98%	100%	102%
A	96%	100%	104%
O	91%	100%	110%
B	61%	68%	111%
I	74%	84%	114%
K	66%	75%	114%
L	75%	90%	120%
C	80%	98%	123%
M	51%	66%	129%
G	57%	90%	158%
F	55%	92%	167%
D	57%	99%	174%

Figure 7-2. Original data on quiz & final scores.

Student	Quiz #1	Final Exam
A	96%	100%
B	61%	68%
C	80%	98%
D	57%	99%
E	90%	82%
F	55%	92%
G	57%	90%
H	98%	82%
I	74%	84%
J	98%	97%
K	66%	75%
L	75%	90%
M	51%	66%
N	74%	60%
O	91%	100%
P	93%	64%
Q	98%	100%

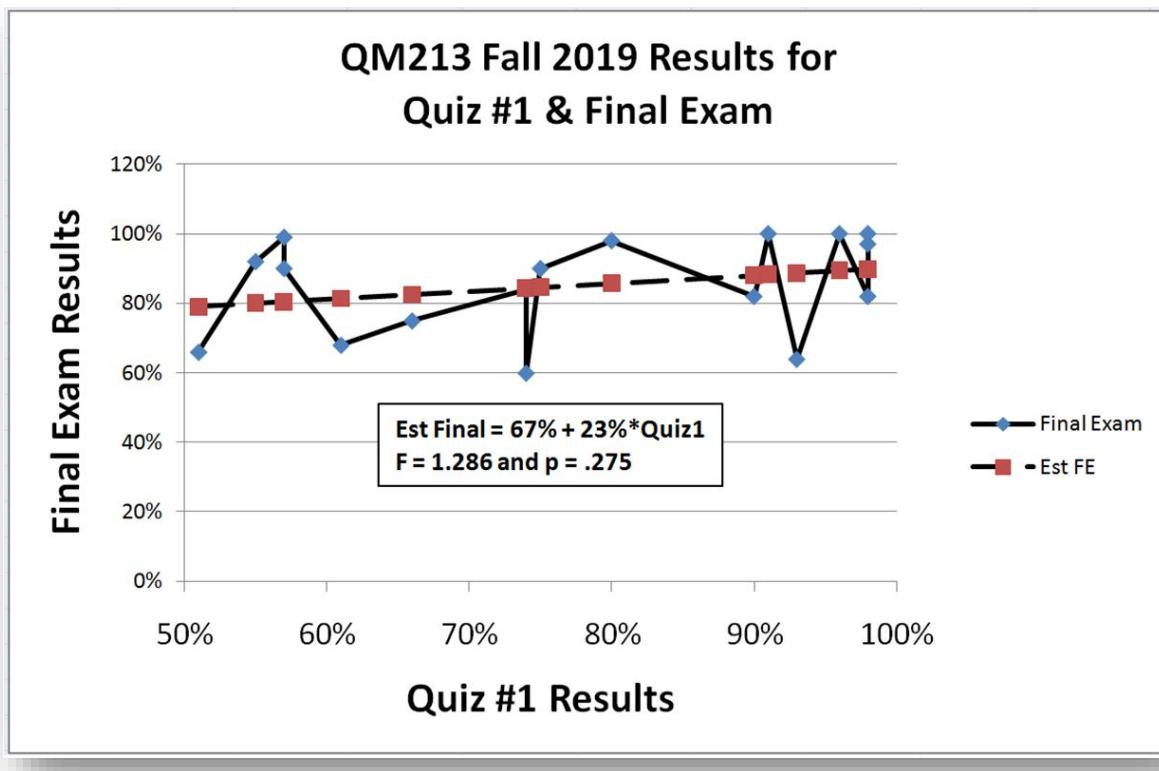
Schlemiel

would leave out students P, N, H, L, C, M, G, F, and D because their results don't match his preconceptions. We call this a *biased* sample. Simply at a gut level, would you trust anything Schlemiel then has to say about the relationship between Quiz #1 results and Final Exam results in QM213 – or in any other population he chose to define for these data?

⁷⁴ A *schlemiel* is a Yiddish term for a dolt.

Let's pursue this example just a bit farther. You will learn later that an *analysis of variance (ANOVA) with regression* could be used to try to predict the Final Exam score based on the collected data about the Quiz #1 results. Without going into detail, the calculations of a regression equation and the ANOVA for the original data are shown in brief in the box in the chart in Figure 7-4. The results don't support the view that there is much of a relation between the first quiz result and the final exam result. The regression equation, such as it is, has a slope (b) of about 23%, implying that the Final Exam score rises by 23% of the Quiz 1 score. But another way of looking at this weak relationship is to examine the coefficient of determination, r^2 (not shown in the figure) which reflects how much of the variability in one variable can be explained by knowing the other. In this case, r^2 turns out to be about 8%; i.e., only 8% of the variability of the Final Exam score can be explained by the Quiz 1 score; all the rest of the scatter has other, unknown sources.

Figure 7-4. ANOVA with regression for unbiased sample.



INSTANT TEST P 7-5

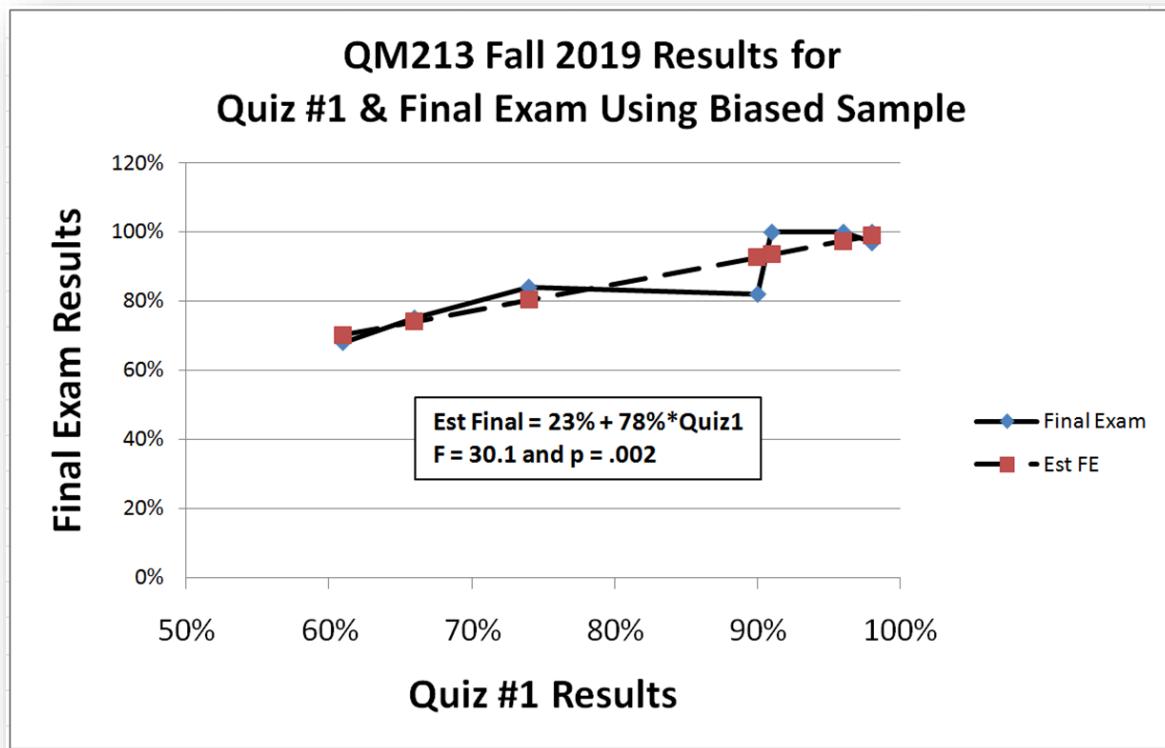
For the diagram above, explain the meaning of the dashed line with the red squares versus the meaning of the blue solid line with the blue diamonds.

Explain why it's OK to use a line graph instead of histograms in the diagram.

Most people would *not* interchange the axes in this graph; they would not put "Final Exam Results" on the abscissa and "Quiz #1 Results" on the ordinate. Why not?

In contrast, the faulty (fraudulent) analysis using Schlemiel's biased sample, shown in Figure 7-5, produces a misleading result exactly in line with Schlemiel's preconceptions: what a surprise! This time the (fake) regression equation has a slope of 78%; the bogus coefficient of determination is a much higher 83%.

So unless someone examines his raw data and catches his deception, Schlemiel can publish his rubbish and [Figure 7-5. Schlemiel's analysis based on biased sample.](#)



distort the literature about the relationship between initial quizzes and final scores. The practical effects, were these lies to become popularly known, might be to discourage students who do poorly in their first quiz in QM213 (the horror!). In fact, the unbiased data do not support such a pessimistic view.

As you can see, defining representative, unbiased samples is critical to honest use of data. The subject of *publication bias*, which is an observed tendency for editors of scientific journals to discourage publication of negative or neutral results, has serious consequences for the ability of researchers to aggregate results from independent studies using the techniques of *meta-analysis* which you will study later in the course.

7.5 Selecting Random Values for an Unbiased Sample

The solution to getting a *representative, unbiased sample* is *random sampling*. In the simplest terms, a random sampling gives every member of a *defined population* an equal chance of being included in the sample being created. In intuitive terms, there are no special rules for selecting the members of the sample – every member of the population might be picked for study.

Looking at the converse, we can say that the population corresponding to a sample is every element which could have been picked according to the definition of the desired population.

Figuring out what the population is for a given sample is not necessarily easy. As a simple example, suppose we want to study the structural resistance to earthquakes of reinforced concrete used in all buildings built in the last decade that use concrete around the world. We take random samples and everything's OK, right? Well no, not necessarily. For one thing, unless we think of it, the sample won't include concrete from buildings that have collapsed in previous earthquakes! So although the population seems at first to be "all buildings on the planet built in the last decade that use reinforced concrete" it's more correctly described as "all buildings on the planet built in the last decade that use reinforced concrete but have not collapsed yet." Don't you think that the sample might be considered biased? After all, the measurements may exclude the buildings that were built in the last decade using shoddy materials such as concrete with much too much sand and too little cement or "reinforced" concrete without metal reinforcement rods.⁷⁵ The information would give a biased view of how strong the concrete actually has been in the last decade.

The easiest way of understanding random sampling is to see it done. Figure 7-6 shows the beginning and end of a long list of 20,000 observations gathered about a total of 1,000 stock brokers, their consumption of alcohol in the hour before a trade (Y/N) and the occurrence of one or more errors in the particular trade (Y/N). How would one select a random sample of 1,000 observations from this list?

One approach (not the only one) is to assign a random number to each observation; in EXCEL, that's easy: the function =RAND() generates a number between 0 and 1 (inclusive) in a uniform distribution. Note that this function takes no argument – the parentheses have nothing between them.

Another EXCEL function is the =RANDBETWEEN(bottom, top) function which generates a uniform distribution of numbers between the limits (inclusive).

The key to these applications for *randomizing data* is that the generated numbers are in uniform distributions, so any number can appear anywhere in the list with equal probability.

Figure 7-6. Start and end of list of 20,000 observations about 1,000 data brokers, their alcohol consumption, and their errors.

		hour before trade	trade
1	123	N	N
2	190	N	N
3	437	N	N
4	614	N	Y
5	197	N	N
6	657	Y	Y
7	286	N	N
8	739	Y	N
9	980	N	N
10	206	Y	N
<hr/>			
19,995	881	N	N
19,996	847	Y	Y
19,997	404	N	N
19,998	896	Y	N
19,999	502	N	N
20,000	457	N	N

⁷⁵ (Associated Press 2011)

We then sort the entire list by the random numbers and pick the first elements in sorted list as our random sample of the desired size. Figure 7-7 shows the results of this process. Note that the middle extract shows the data around the desired limit of 1,000 entries. We have but to select the first 1,000 entries in the sorted list to have a random sample from the entire 20,000 of the original data.

Figure 7-7. Original data with random numbers assigned and used to sort the data.

Sequence #	Random #	Obs #	Broker	Alcohol consumed 1 hour before trade	Errors in trade
1	0.00181	10,600	140	N	Y
2	0.00461	14,956	245	N	N
3	0.02459	13,649	608	N	N
4	0.03492	19,998	896	Y	N
5	0.04155	8,326	994	N	Y
<hr/>					
998	0.1030	6,943	491	N	Y
999	0.1032	18,171	151	N	Y
1,000	0.1053	2,160	986	Y	N
1,001	0.1053	7,217	467	N	Y
1,002	0.1075	19,235	959	N	N
1,003	0.1084	16,737	471	Y	Y
<hr/>					
19,995	0.99136	1,799	639	Y	Y
19,996	0.99325	1,505	36	N	N
19,997	0.99401	1,599	467	Y	N
19,998	0.99429	2,348	856	N	N
19,999	0.99446	8,069	17	N	Y
20,000	0.99903	6,317	812	N	N

Looking at this procedure step by step,

- We start by assigning a random number to each of the observations using the =RAND() function in EXCEL.
- Once we've created the list of 20,000 random numbers, we fix them (stop them from changing) by copying the entire list and pasting it as values into the first cell of the list, freezing the numbers as values instead of functions..
- Finally, we sort the list using the random numbers, as shown in Figure 7-7.
- We select the first 1,000 rows and that's it: a random sample of 1,000 records from the original 20,000.
- There is no human judgement (and potential bias) is involved in the choice. This method is easy to apply to any data set and always works.

INSTANT TEST P 7-8

Create a list of 20 random values from 5,000 to 10,000 using
 $=INT(NORM.INV(RAND(),5000,10000))$. Copy the list and Paste Special into another column using the Values option to freeze the data. Now apply random numbers using =RAND() and then copy/paste-special next to your 100 values. Sort the two columns by the RAND() values and practice selecting random samples of size 20 from the list.

7.6 More about Probability and Randomness

You've heard about and perhaps even studied probability in other courses; in this course, we've already introduced some basic ideas about probability in §5.3. For now, it suffices to establish probability as a measure of what we expect in the long run: what happens on average when we look at repeated actions in a defined situation.

In §7.4 on random sampling, we used the =RAND() function of EXCEL. EXCEL's HELP function describes RAND() as follows: "Returns an evenly distributed random real number greater than or equal to 0 and less than 1." The phrase *evenly distributed* refers to what mathematicians and statisticians refer to as the uniform probability distribution (§5.4). For the RAND() function, we can assert that the frequency of occurrence of numbers 0, 0.1, 0.2... 0.8, and 0.9 are all equal (to 10% of the observations) on average over the long run. But we can also assert that the occurrence of numbers 0, 0.01, 0.02, 0.03... 0.96, 0.97, 0.98 and 0.99 are also equal (to 1% of the observations) on average over the long run. The generated numbers are random precisely because there is equal frequency of all the numbers regardless of precision (within the limits of calculation of EXCEL).

Because the numbers are generated by mathematical processes that actually generate exactly the same sequence of numbers if they start from any given starting point, we call the =RAND() function a *pseudo-random number generator*. In other words, the output looks random even though the sequence is theoretically repeatable.

But wouldn't a generator that produced the sequence 0.1, 0.2, 0.3, 0.4... and so on in perpetuity, in the same order, produce equal frequencies of numbers of whatever precision we chose? Yes, but they wouldn't be considered random. Randomness applies also to the *sequence* of the data. There must be no predictability of which number follows any given number in a random sequence. So the frequency of, say, the digit 1 followed by the digit 2 or digit 1 followed by digit 3 must be equal, and so on.

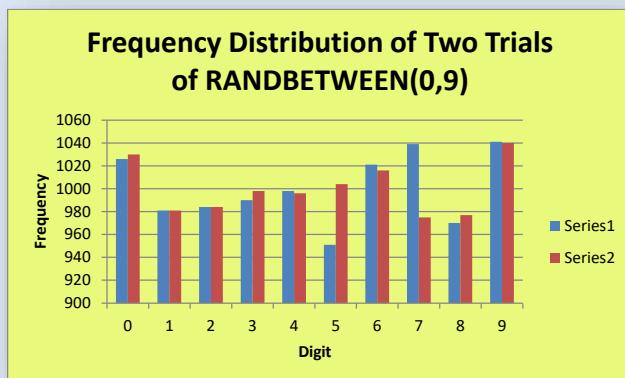
Another function that creates pseudo-random numbers in EXCEL is =RANDBETWEEN(bottom,top). For example, =RANDBETWEEN(0,9) might produce the sequence 5 1 0 8 1 5 2 9 7 3 in one usage and 1 8 5 4 6 3 2 9 0 7 in the next.

Later we will study methods of applying statistical inference to this kind of analysis: we will be using Goodness-of-fit tests to evaluate the likelihood that a set of frequency observations deviate from expectation by random sampling alone.

INSTANT TEST P 7-9

Use Excel to create a list of 10,000 pseudo-random whole numbers between 0 and 9. Generate a frequency distribution showing the frequency of each digit in the list. Then do it again and compare the results. Are they what you expected?

Here are the results of such an exercise carried out by the author:



7.7 Random Number Generators

Some mathematical functions have been written that appear to create random number sequences. A simple one is the series of digits in the decimal portion of the number π ; another one is a simple procedure involving taking the fractional part of a starting number, using it as the exponent of a calculation, and taking the fractional part of the result as the next “random” number in the series – and then starting over using this new “random” number as the exponent for the next step. All these methods are called *iterative pseudo-random number generators* because they produce sequences that superficially look random but which in fact are perfectly repeatable and predictable if you know the rule (the *algorithm*) and the starting point (the *seed value*).

Another problem with iterative pseudo-random number generators is the precision of the calculations: eventually, it is possible for a generated number to be created that has already occurred in the sequence. The moment that happens, the sequence enters a loop. For example, if we considered the reduction to absurdity of having an iterative pseudo-random number generator that truncated its calculations at a single decimal digit, then the only numbers it could generate would be 0, .1, .2, .3... and .9. Suppose the sequence it generates were .4, .2, .5, .6 and then .4 again: the system would enter the endless loop of .4, .2, .5, .6, .4, .2, .5, .6, .4, .2, .5, .6 and so on. Not very random, eh?

The point of raising these issues here is not to make you experts on random number generators: it's to make you think about the concept of randomness and the fundamentals of probability. For now, it's enough to have you thinking about probability as the *expectation of observations in a random process*; that is, as average frequencies of occurrence for processes that have no obvious pattern.

7.8 Probabilities in Tossing Coins

A classic example used in introducing probabilities is the tossing of coins. A coin is tossed and lands with either one side (heads) or the other (tails) facing up. We deliberately ignore the rare cases where the coin lands on its edge and stays that way. We say that the probability of heads is $\frac{1}{2}$ and the probability of tails is $\frac{1}{2}$.

In general, the sum of the probabilities of all possible results in a defined system is always 1. The probability of impossible events is 0. So the probability of heads and the probability of tails in a single coin-toss is 1. The probability of 2 heads plus the probability of 1 head and 1 tail plus the probability of 2 tails in tossing a coin twice (or tossing two coins at the same time) is 1. We could write this latter assertion as

$$P\{H,H\} + P\{H,T\} + P\{T,H\} + P\{T,T\} = 1$$

7.9 Probabilities in Statistical Inference

In the ANOVA tables you have seen in previous sections and on the regression charts there were figures labeled p . These refer to the probability that there is no relationship among the data analyzed (that idea is called the *null hypothesis*); it is a measure of how likely we are to see results as deviant from the most likely expected result or more deviant by pure luck – by chance alone. As you will see in the discussion of hypothesis testing, looking at how likely our observed results are as a function of chance variations is a core idea for modern statistics.

INSTANT TEST P 7-10

Explain to yourself or to a buddy why the probability of getting two heads on top if you toss two coins is $\frac{1}{4}$. Then explain why the probability of getting one head and one tail on top is $\frac{1}{2}$ instead of $\frac{1}{4}$.

7.10 The Central Limit Theorem in Practice

What happens when we sample from a population?

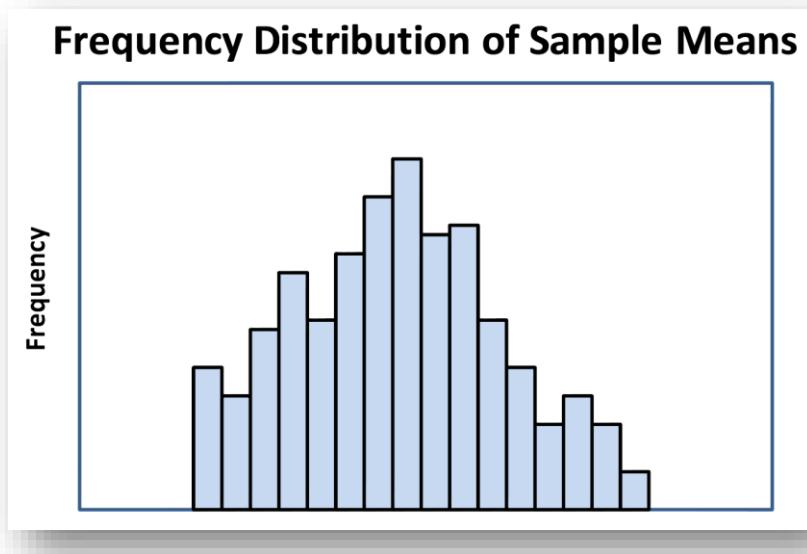
Does the sample reflect the characteristics of the population? Yes, but not in the sense of matching it exactly. Suppose we have a population of 4,821 widgets produced from assembly line #3 at the Urgonian Corporation plant in Olympus Mons on June 14, 2219. The parametric mean length of the widgets is determined to be exactly 342 mm by measuring every single widget; the parametric standard deviation of the length is exactly 0.716 mm.

But now we take a sample of 100 widgets from the batch of 4,821 and discover that the sample mean of the lengths is 344 mm and the standard deviation is 0.922 mm. Then we take another sample of 100 widgets and – horrors – it doesn't match the population either: the mean length is 341 mm and the standard deviation is 0.855 mm.

There's nothing wrong here. We are seeing a demonstration of sampling variability and of the Central Limit Theorem. The interesting and important aspect of sampling is that, according to the Central Limit Theorem, the more samples we take, the closer the overall average of the sample statistics approaches the parametric value.

As we accumulate data from dozens of samples, we can build a frequency distribution showing how often different means occur in the samples; Figure 7-8 shows what such a graph might look like.

Figure 7-8. Means from dozens of samples.



As the number of samples grows, what we find is that

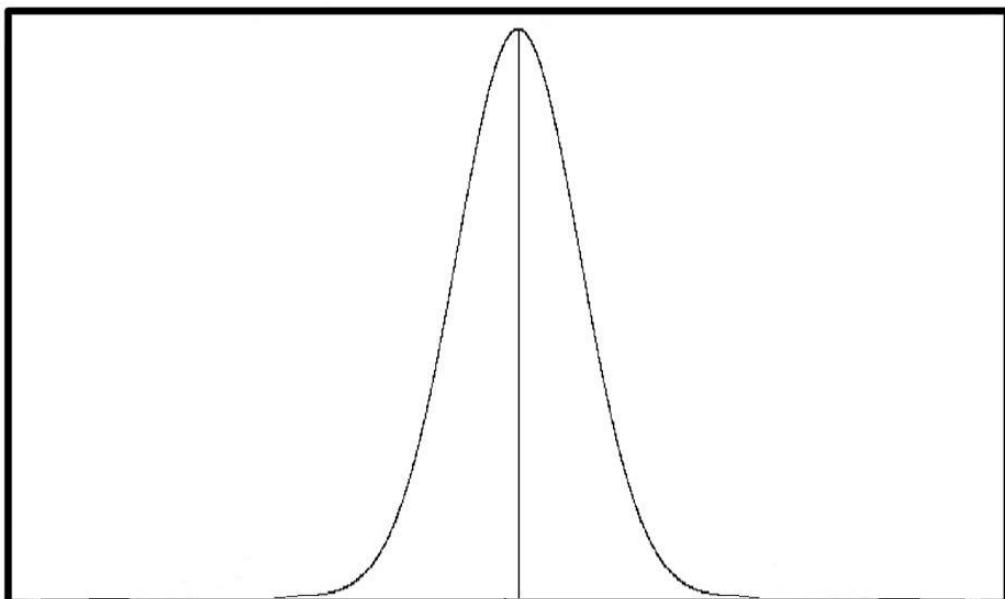
- The distribution curve becomes more and more symmetrical;
- The curve gets smoother-looking;
- The mean of the distribution (\bar{Y}) approaches the parametric mean (μ) more and more closely;
- The variance (and therefore standard deviation) of the distribution gets smaller and the curve gets tighter around the mean.

Figure 7-9 shows the results of these tendencies as the number of samples grows into the hundreds. The image is inserted full width on the page because careful examination will reveal that the curve is actually a step function corresponding to the hundreds of samples. If there were thousands of samples, the curve would look smooth at this scale and would be even narrower around the mean.

Figure 7-9. Sampling distribution with hundreds of samples.

Frequency Distribution of Sample Means

Frequency



Sample Mean

The effect of the Central Limit Theorem is stronger as the size of the individual samples rises; samples of size 100 show a faster approach to the kind of distribution shown in than samples of size 10.

The distribution shown in Figure 7-9 is a Normal distribution. The Central Limit Theorem can be stated in intuitive terms as follows:

As sample size increases, the means of random samples drawn from a population of *any* underlying frequency distribution will approach a Normal distribution *with its mean corresponding to the parametric mean of the source distribution*.

The Central Limit Theorem is enormously important in applied statistics. It means that even if an underlying phenomenon doesn't show the attributes of the Normal distribution, the means of samples will be normally distributed. Since so much of modern statistics assumes a Normal distribution for the underlying variability of the phenomena being analyzed, the Central Limit Theorem means that we can circumvent non-normality by working with samples of observations – groups of data – instead of with individual observations.

7.11 The Expected Value

The Central Limit Theorem also brings to light another concept of great importance: the *expected value* of a statistic. The expected value is the average of a statistic computed over an infinite number of samples.

For example, the *expected value of the sample mean* is the *parametric mean*, μ . We say that the observed sample mean is a *point estimator* of the parametric mean – a single value that indicates what the parameter may be.

Statisticians have also shown that the expected value of the variance of the sample means ($\sigma^2_{\bar{Y}}$) is the ratio of the parametric variance (σ^2) to the sample size (n) and thus the expected standard deviation of the sample means ($\sigma_{\bar{Y}}$) is expected to be the parametric standard deviation (σ) divided by the square root of the sample size(\sqrt{n}):

$$\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n} \quad \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of the mean is called the *standard error of the mean*. In general, the standard deviation of any statistic is known as the *standard error* of that statistic.

Another interesting application of the Central Limit Theorem is that, in the absence of any further information, whatever we encounter is most likely to be *average*. So for example, suppose we are working on a very important project such as writing a statistics textbook and the phone rings; the caller-ID shows an unknown caller. In the absence of further information, the call is most likely to be of *average importance*.⁷⁶

Therefore, one can reasonably defer answering the call and allow it to go to voice-mail for later processing.

Similarly, if one's activity is less than average in importance (for instance, watching the Olympic Paint Drying Championships on holovision), then one can reasonably answer an unknown caller.

Statistics in action!

7.12 More About the Normal Distribution

In this text, the *Normal distribution* has a capital N for *Normal* to be sure that no one thinks that there is anything abnormal about non- Normal distributions! The Normal distribution plays an important role in statistics because many ways of describing data and their relationships depend on a Normal error distribution. For example, in the linear regression that you will study later, the error term ε in the equation

$$Y_{ij} = a + bX_i + \varepsilon_{ij}$$

represents a Normally-distributed error with a mean of zero and a variance defined by what is called the *Residual MS (residual mean square)* in the ANOVA table. In other words, the linear model defines a best-fit line for Y, the expected or predicted dependent variable, as a function of the Y-intercept (a , the value of Y when X, the independent variable, is zero) plus the product of the slope b times the value of X, plus the random (unexplained) error ε .⁷⁷

Most of the descriptive statistics and statistical tests we use routinely are called *parametric statistics* because they assume a Normal distribution for the error term or unexplained variance. ANOVA, ANOVA with regression, t-tests, product-moment correlation coefficients (all to be studied in detail later in this course) uniformly assume a Normal error distribution. There are other assumptions too, which we will also discuss later; one important concept is that the mean and the variance of a statistic are supposed to be *independent*. That is, for

⁷⁶ ...and average duration and average volume and average origin and average annoyance-value and average....

⁷⁷ As explained in the Preface, most textbooks *never* make forward references. However, this textbook often makes such references so that when students come to study the details of a technique, they will have encountered it – and some modest amount of information about how it is used – long before they have to learn the details.

parametric statistical analysis like t-tests and ANOVA, we assume that small observed values have the same variability as large observed values for the same variable.

A counter-example is the relation between, say, size and weight. No one would expect the variance of the weights of tiny model sailboats weighing a kilogram or so to be the same as the variance of the weights of battleships weighing tens of thousands of tons.

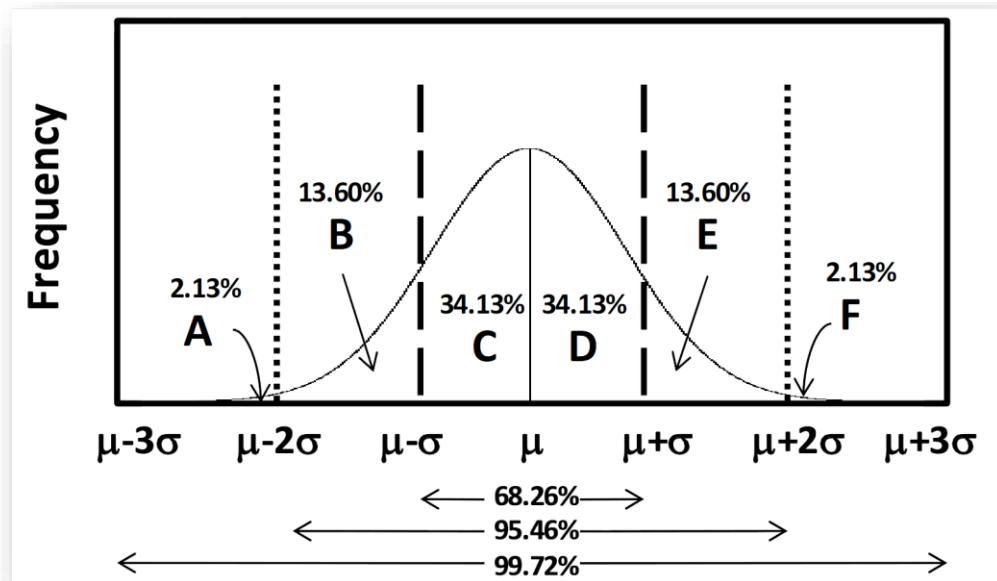
When these assumptions are not justified, we have to use *non-parametric statistics*. Examples include using the median and the mode instead of the mean and the range and other non-parametric measures of variability instead of the variance or standard deviation. Non-parametric tests can include comparisons of ranks instead of measured variables. Examples of such tests include

- Kruskal-Wallis Test for comparisons of central tendency of ranked data
- Friedman's Method for Randomized Blocks for comparisons of data that don't fit the Normal curve
- Mann-Whitney U-Test for comparing percentages
- Kolmogorov-Smirnov Two-Sample Test for comparing frequency distributions
- Wilcoxon's Signed Ranks Test for Two Groups of Paired Data
- Spearman's Coefficient of Rank Correlation.⁷⁸

We use the Normal distribution so much that some attributes become familiar simply by force of repetition.⁷⁹

Figure 7-10 shows some of the most often used characteristics of the Normal distribution: the areas under the curve as a function of distance from the mean (μ) measured in standard deviations (σ).

Figure 7-10. Characteristics of the Normal distribution.



⁷⁸ The names of these tests are included in case you desperately need to analyze rank data or other non-Normal data; you can look up how to perform the tests in your statistical package or in standard textbooks. Eventually, this text will expand to include such methods for upper-year courses.

⁷⁹ Don't memorize this stuff – just learn it by using it!

In Figure 7-10, the letters and numbers indicate the following relationships:

- The probability that an observation picked at random from a Normally-distributed variable with defined mean μ and standard deviation σ will be smaller than three standard deviations (often spoken of as *less than three sigmas*) away from the mean is about 2.13%. This area corresponds to the section marked A in the figure. We can write this statement conventionally as

$$P\{Y \leq \mu - 3\sigma\} \approx 2.13\%$$

- In other words, the probability of encountering a normal variate that is at or less than 3 sigmas below the parametric mean is 2.13% or roughly 1 in 47 tries.
- Because the Normal distribution is perfectly symmetric, area F is identical to area A; that is,

$$P\{Y \geq \mu + 3\sigma\} \approx 2.13\%$$

- Section B in the figure represents the 13.6% probability that an observation picked at random from a Normally distributed variable will lie between 1 and 2 sigmas below the mean. That is,

$$P\{ \mu - 2\sigma \leq Y \leq \mu - \sigma \} \approx 13.6\%$$

- Section E corresponds to section B on the other side of the mean, so we can also write

$$P\{ \mu + \sigma \leq Y \leq \mu + 2\sigma \} \approx 13.6\%$$

- Areas C and D correspond to the chance of picking a variate at random which lies within one sigma of the mean. Together, C and D add up to 68.2% of the total area under the curve (shown on the arrows below the abscissa), implying that more than 2/3 of all values picked at random from a Normally distributed population will lie between $\mu - \sigma$ and $\mu + \sigma$.

$$P\{ \mu - \sigma \leq Y \leq \mu + \sigma \} > 67\%$$

- Similarly, the figure illustrates the rule of thumb that about 95% (95.46%) of randomly-chosen members of a Normally-distributed population will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$ (that is, within two sigmas of the mean).

$$P\{ \mu - 2\sigma \leq Y \leq \mu + 2\sigma \} > 95\%$$

- More than 99% (99.72%) of the population lies between $\mu - 3\sigma$ and $\mu + 3\sigma$ (within three sigmas of the mean).

$$P\{ \mu - 2\sigma \leq Y \leq \mu + 2\sigma \} > 99\%$$

In ordinary life, we can make use of these approximate values when evaluating statements about the Normality or non-Normality of specific observations if we know something about the underlying population distribution and have some confidence that the error distribution is Normally distributed.

For example,

- Imagine that we know that the average height of Lunar Colony men of 19 years of age in a study is 172.1 cm with standard deviation of 2.8 cm.
- We can assert from these data that about 2/3 of the 19-year-old males in the Lunar Colony have a height between $172.1 - 2.8$ cm and $172.1 + 2.8$ cm.
- That's about 66% of the 19-year-old boys between 169.3 cm and 174.9 cm.
- For you non-metric folks, that's about 6' 3.6" and 6' 6.1" with a mean of 6' 4".
- Similarly, about 99% would be within three sigmas, which would be 163.7 cm and 180.5 cm (6' 1.1" and 6' 8.6").

- So if someone said of a young man, “Wow, he’s really exceptionally short” because he was only 6’ 2” high, we could identify the statement as inaccurate – unless the speaker’s definition of “exceptionally” were unusually liberal and included men within the range of 99% of the Lunar Colony population of that age.

7.13 Statistical Inference: Interval Estimation

Knowing that random samples vary according to particular patterns – for instance, the means of samples approach a Normal distribution – means that we can *estimate the parametric value* based on a *sample value*.

For example, if we sample randomly from the Lunar Colony boys of 19 earth years and measure the heights of 25 of them, the mean of that sample should tell us something about the mean of the population. Using the Central Limit Theorem, we assert that *our best estimate of the parametric mean*, in the absence of any other information, *is the sample mean*.

However, common sense tells us that the sample mean of any one sample may differ from the parametric mean; intuitively, it doesn’t seem reasonable to expect that a random sample would magically be exactly centered on the population mean. Therefore, we compute an *interval estimate* for the parametric mean using our knowledge of the sample mean and of the variability and pattern of distribution of such means.

An *interval estimate* for any statistic is a range with lower and upper *confidence limits*. Typical $(1 - \alpha)$ confidence limits for any interval estimate of a parametric value are called the $(1 - \alpha)$ confidence limits. For example, we often refer to the 95% *confidence limits* of a statistic, where $\alpha = 0.05$. Another common choice is the 99% confidence limits of a statistic, where $\alpha = 0.01$.

These intervals are interpreted as follows:

- The *probability of being correct* in asserting that the $(1 - \alpha)$ confidence limits include the value of the parametric statistic is $(1 - \alpha)$.
- The *probability of being wrong* in asserting that the $(1 - \alpha)$ confidence limits include the value of the parametric statistic is α .

Here are some examples of interval estimates for a variety of made-up statistics and different ways of interpreting them:

- The sample mean cost of a trans-Jovian flight in 2219 is 1,452 credits; the 95% confidence limits are 1167 and 1736 credits. There is a 95% chance of being correct in guessing that the mean cost lies between 1167 and 1736 credits. There is therefore a 5% chance of being wrong in that assertion.
- A sample of Martian fornselling beans has a mean growth potential of 182% per month; the 90% confidence limits are 140% to 224%. There is only a 10% chance of being wrong in claiming that the growth potential is between 140% and 224% per month.
- A study of Norwich University students’ usage of fornselling chips consumed per month in 2219 showed an average of 3.8 kg per student with 80% confidence limits of 2.3 to 5.3 kg. We would be right 80% of the time that we repeat this kind of sampling and computation of the confidence limits for the consumption figures. We’d be wrong in 20% of the estimates based on samples of that size.
- The Gorgonian factor calculated for a sample of 3,491 Sutellian customers indicated an average time to immobility upon exposure to the Gorgonian advertisements of 2 hours 12 minutes with 99% confidence limits of 1 hour 54 minutes through 2 hours 36 minutes. Our chance of being wrong in using this procedure to guess at the correct time to immobility is only 1 time out of a hundred. That is, if we were to repeatedly take random samples of size 3,491 Sutellians exposed to the same ads, in 99 out of a hundred experiments, the computed interval estimates would correctly include the parametric mean time to immobility.
- The mean variance for the sales resulting from exposure to a modified mind-control regimen projected through the holographic broadcasting networks was 3,622 with 95% confidence limits of

1,865 and 5,957. Using this confidence-interval calculation procedure, we have a 95% probability of really including the true parametric sales figure in our computed interval.

Students may have noticed that in *no case* above were the confidence intervals interpreted as follows: “*The probability that the parametric statistic is between the lower and upper (1 - α) confidence limits is (1 - α).*” All of the interpretations were in terms of the chance of *being right (or wrong)* in asserting that *the limits include* the parametric value, *not* the probability that the parameter is between the limits. The parameter is fixed for a population; it is the estimates that vary around it. Sokal and Rohlf explained this subtle point as follows in their classic textbook:

“We must guard against a common mistake in expressing the meaning of the confidence limits of a statistic. When we have set lower and upper limits ... to a statistic, we imply that the probability of this interval covering the mean is 0.95, or, expressed in another way, that on the average 95 out of 100 confidence intervals similarly obtained would cover the mean. We cannot state that there is a probability of 0.95 that the true mean is contained within any particular observed confidence limits, although this may seem to be saying the same thing. The latter statement is incorrect because the true mean is a parameter; hence it is a fixed value and it is therefore either inside the interval or outside it.”

It cannot be inside a particular interval 95% of the time. It is important, therefore, to learn the correct statement and meaning of confidence limits.”⁸⁰

Notice that most confidence limits are symmetrical, but that the mind-control variance was not; there is no guarantee that confidence limits will be symmetrical. The exact calculations for the upper and lower limits depend on the nature of the variation of the sample statistics. For example, most statistics are normally distributed, but percentages in the low (less than 10%) and high (greater than 90%) ranges are typically not normally distributed because there is greater uncertainty (and therefore variability) about the extremes than about the central portion of the distribution. We will learn more about how to handle such non-Normal measurement scales in the introduction to *data transforms* later in the text.

7.14 Population Mean Estimated Using Parametric Standard Deviation

One of the most common calculations in statistics is the estimation of the confidence limits for a mean. The exact calculations depend on whether we know the parametric standard deviation or not.

When we know the population standard deviation (for example, if the population has long been studied and the variance of the statistic in question is established from many repeated measurements and no longer considered to be an estimate) then we can use it directly in the calculation of the confidence limits.

In our introduction to the Normal distribution, we learned that for a Normally distributed variable with mean μ and standard deviation σ , 95% of the values lie between -1.96σ and $+1.96\sigma$ from the mean, μ . That is,

$$P = \left\{ -1.96 \leq \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \leq +1.96 \right\} = 0.95$$

Recalling that for samples of size n , sample means \bar{Y} are normally distributed around the parametric mean with standard error of the mean

$$\sigma_{\bar{Y}} = \sigma / \sqrt{n}$$

we can thus write the calculation for the 95% confidence limits to the mean as

$$P\{\bar{Y} - (1.96\sigma / \sqrt{n}) \leq \mu \leq \bar{Y} + (1.96\sigma / \sqrt{n})\} = 0.95$$

Or more generally, where z_α is the critical z-score corresponding to the probability α to the left of its value, then correspondance

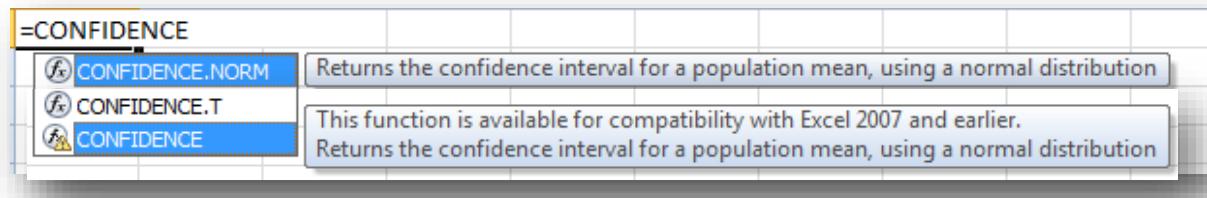
⁸⁰ (Sokal and Rohlf, Biometry: The Principles and Practice of Statistics in Biological Research 1981) p 144.

$$P\{\bar{Y} - z_{\alpha/2} s_{\bar{Y}} \leq \mu \leq \bar{Y} + z_{\alpha/2} s_{\bar{Y}}\} = 1 - \alpha$$

The composite image in Figure 7-11 shows the two EXCEL 2010 functions that perform this calculation automatically:

Unfortunately, the **HELP** text in EXCEL 2010 for these functions has a mistake, as highlighted by the red underline in the composite image below (Figure 7-12).

Figure 7-11. Calculating confidence limits for the mean when the *parametric* standard deviation is known.



underline in the composite image below (Figure 7-12).

This function computes *half* of the confidence interval, not the confidence interval or the confidence limits.

Figure 7-12. HELP text with error.

The formula shown at the lower left of Figure 7-12 defines the upper and lower confidence *limits*, not the confidence *interval*. The \pm symbol implicitly defines (using the $-$) the *lower confidence limit* (sometimes denoted L_1) and (using the $+$) the *upper confidence limit*, sometimes denoted L_2 . The confidence *interval* is $L_2 - L_1 = 2 * 1.96(\sigma/\sqrt{n})$.

Using the function name and our \bar{Y} variable,

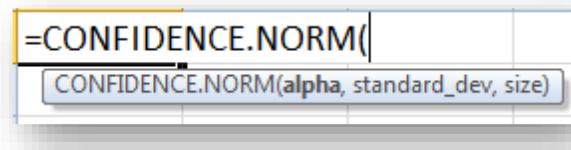
$$L_1 = \bar{Y} - \text{CONFIDENCE.NORM}(\text{parms}) \text{ and}$$

$$L_2 = \bar{Y} + \text{CONFIDENCE.NORM}(\text{parms}).$$

Figure 7-13 shows the parameters (*parms*) for =CONFIDENCE.NORM.

- The parameter **alpha** is the complement of the confidence level; thus for 95% confidence, $\alpha = 0.05$.
- The parameter **standard_dev** is the parametric standard deviation.
- The parameter **size** is the sample size n .

For example, if we know from [Figure 7-13. Parameters for computing confidence limits to the mean given the f parametric standard deviation.](#)



Barsoomian roncinal weights is 23 kg and we acquire a sample of 12 roncinals whose mean weight is 745 kg, we can easily calculate that the lower and upper confidence 95% confidence limits are as shown in the composite image of Figure 7-14.

Figure 7-14. Calculating lower and upper confidence limits for the parametric mean given the parametric standard deviation.

	A	B		A	B
1	Sample mean:	745	1	Sample mean:	745
2	Confidence level:	95%	2	Confidence level:	0.95
3	Parametric std dev:	23	3	Parametric std dev:	23
4	Sample size:	12	4	Sample size:	12
5	Function output:	13.013	5	Function output:	=CONFIDENCE.NORM((1-B2),B3,B4)
6	Lower confidence limit:	732.0	6	Lower confidence limit:	=\$B\$1-\$B\$5
7	Upper confidence limit:	758.0	7	Upper confidence limit:	=\$B\$1+\$B\$5

INSTANT TEST P 7-19

Duplicate the calculations shown above in your own spreadsheet but don't use any \$ signs in the formulas so you can propagate the formulas sideways. Create 6 columns of data with confidence levels 80%, 85%, 90%, 95%, 99% and 99.9%. Graph the lower and upper confidence limits against the confidence level. Discuss your findings in the discussion group on NUoodle for this week.

7.15 Estimating Parametric Mean Using the Sample Standard Deviation

What happens if we don't know the *parametric* standard deviation (which is the same as saying we don't know the parametric variance)?

We use the *sample* standard deviation, s to compute an *estimate* of the standard error of the mean

$$s_{\bar{Y}} = s/\sqrt{n}$$

where n is the sample size and s is the standard deviation of the sample. Then the statistic

$$(\bar{Y} - \mu)/s_{\bar{Y}}$$

is distributed as a *Student's-t distribution* with $n - 1$ degrees of freedom.

These deviates will be more variable than those computed using a single parametric value for the standard error because sometimes the sample s will be smaller than the parametric σ and sometimes it will be larger. It follows that the frequency distribution for the ratio

$$(\bar{Y} - \mu)/s_{\bar{Y}}$$

must be different from the distribution of

$$(\bar{Y} - \mu)/\sigma_{\bar{Y}}$$

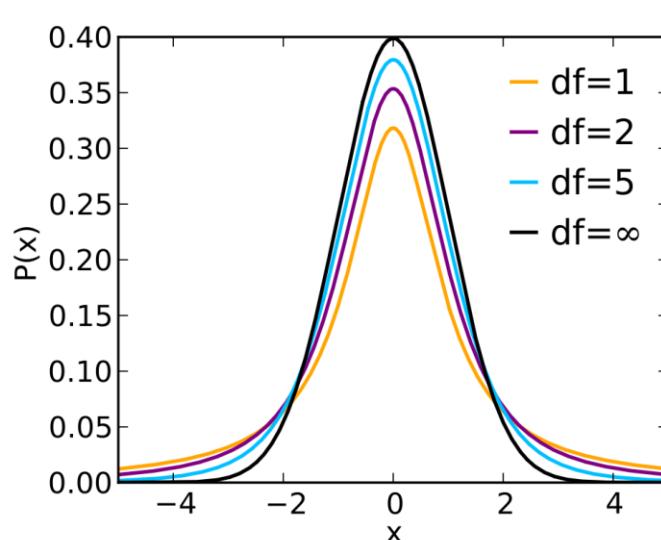
that we studied earlier: it will be broader because of the variations in the denominator.

In fact the distribution of

$$(\bar{Y} - \mu)/s_{\bar{Y}}$$

is called *Student's-t distribution* and was published in 1908 by the famous English mathematician William Sealy Gosset (1876-1937) who published extensively in statistics under the pseudonym *Student*. The distribution is actually a family of curves defined by the sample size: the *degrees of freedom* of each distribution is one less than the sample size on which the sample standard deviation is computed. Note that when $df = \infty$, Student's-t distribution is the Normal distribution. Figure 7-15 illustrates this relationship between Student's t and the Normal distribution.⁸¹

Figure 7-15. Family of Student's-t distributions.



⁸¹ Image used in compliance with Creative Commons Attribution 3.0 license from <http://upload.wikimedia.org/wikipedia/commons/thumb/4/41/Student_t_pdf.svg/1000px-Student_t_pdf.svg.png> or <<http://tinyurl.com/9abxycu>>.

7.16 Degrees of Freedom Vary in Statistical Applications

We use *degrees of freedom* (df) extensively in our work in statistics. One interpretation is that if we have n data in a sample, calculating the *sum of the values* fixes $(n - 1)$ of the values; i.e., knowing the sum, we don't need to know the last of the values, since it can be computed as the sum minus the sum of the other $(n - 1)$ data. Thus only $(n - 1)$ of the data are free to vary – hence the degrees of freedom are $(n - 1)$. However, the exact computation of the degrees of freedom for a statistic is particular to each type of statistic.

As mentioned in the previous section, Student's-t distribution approaches the Normal distribution more and more closely as the degrees of freedom rise; indeed the Normal distribution *is* Student's-t distribution with infinite degrees of freedom. Think of the approach of Student's-t distribution to the Normal distribution with increasing sample size as another example of the Central Limit Theorem.

7.17 Notation for Critical Values

The *critical value* of Student's-t distribution with $n - 1$ degrees of freedom below which α of the distribution lies is written as

$$t_{\alpha[n-1]}.$$

In more general use, the degrees of freedom are represented by the letter v (*ν*), the Greek equivalent of n . Thus you might see this form:

$$t_{\alpha[v]}$$

to represent the critical value of Student's t for v degrees of freedom which has a probability of α of having values that large or smaller. On the probability distribution, $t_{\alpha[v]}$ thus demarcates the portion α of the curve to the left and the portion $(1 - \alpha)$ to the right of that value. The square brackets are a typical way of separating the degrees of freedom from the critical probability.

7.18 Two-Tailed Distributions

Because the Normal distribution and the Student's-t distribution are symmetric around the mean, they are called *two-tailed* probability distributions.

In practice, we express the confidence limits based on a sample mean \bar{Y} with unknown parametric standard deviation as follows:

$$P\{\bar{Y} - t_{\alpha/2[n-1]}S_{\bar{Y}} \leq \mu \leq \bar{Y} + t_{\alpha/2[n-1]}S_{\bar{Y}}\} = 1 - \alpha$$

Notice that α represents the *total of the area* below and above the $(1 - \alpha)$ confidence limits. Each *tail* of the distribution represents a probability of $\alpha/2$.

So to compute the $(1 - \alpha)$ confidence limits of a population mean given the sample mean \bar{Y} and that sample's standard deviation s and sample size n , we

- (1) Compute the standard error of the mean as

$$S_{\bar{Y}} = s/\sqrt{n}$$

- (2) Locate the absolute value (e.g., $| -3 | = 3 = | +3 |$) of the t-statistic corresponding to a *left tail* of probability $\alpha/2$; that is,

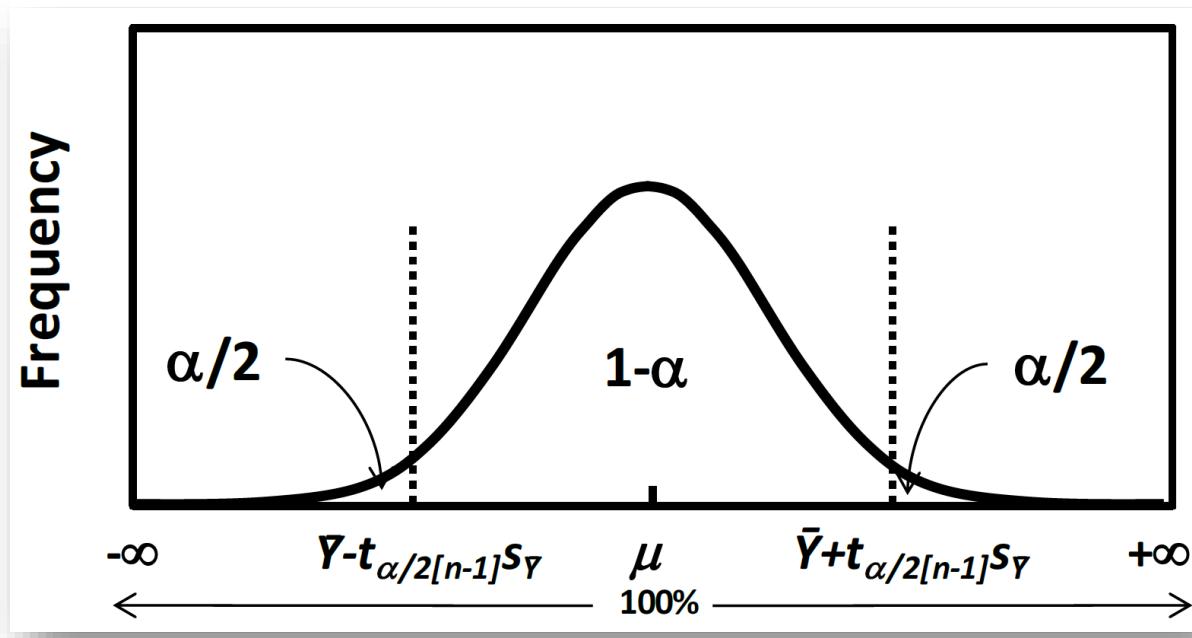
$$| t_{\alpha/2[n-1]} |$$

- (3) Compute the lower and upper confidence limits as

$$L_1 = \bar{Y} - | t_{\alpha/2[n-1]} | S_{\bar{Y}} \quad \text{and} \quad L_2 = \bar{Y} + | t_{\alpha/2[n-1]} | S_{\bar{Y}}$$

Figure 7-16 shows the two-tailed probabilities for distributions like the Normal and the Students-t.

Figure 7-16. Confidence limits for the mean using Student's-t distribution and two-tailed probabilities.



7.19 EXCEL CONFIDENCE.T Function

The EXCEL 2010 function, =CONFIDENCE.T, that computes half the confidence interval for a parametric mean based on a sample mean \bar{Y} and its observed sample standard deviation s . The function is illustrated in the center of Figure 7-11 and is highlighted below in Figure 7-17.

Figure 7-17. Function for confidence limits of a mean knowing the *sample* standard deviation.

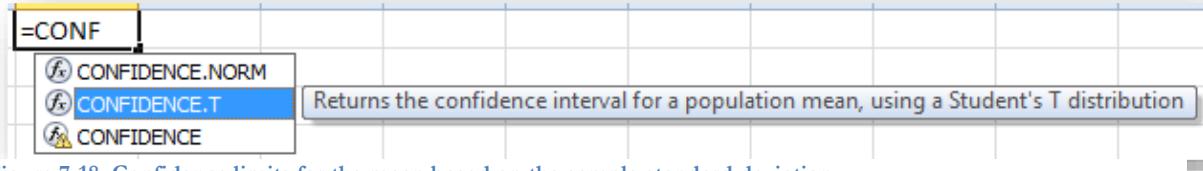


Figure 7-18. Confidence limits for the mean based on the sample standard deviation.

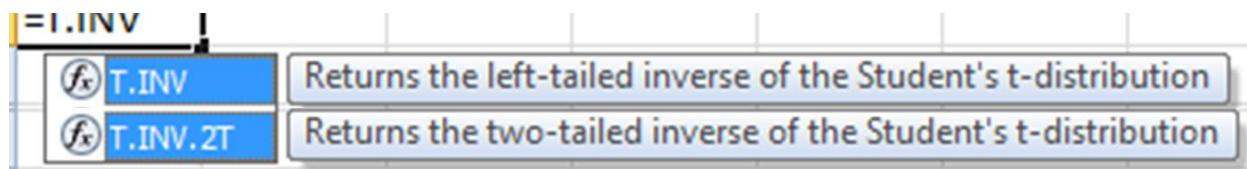
	A	B		A	B
1	Sample mean:	745	1	Sample mean:	745
2	Confidence level:	95%	2	Confidence level:	0.95
3	Sample std dev:	23	3	Sample std dev:	23
4	Sample size:	12	4	Sample size:	12
5	Function output:	14.614	5	Function output:	=CONFIDENCE.T((1-B2),B3,B4)
6	Lower confidence limit:	730.4	6	Lower confidence limit:	=\$B\$1-\$B\$5
7	Upper confidence limit:	759.6	7	Upper confidence limit:	=\$B\$1+\$B\$5

=CONFIDENCE.T works much like =CONFIDENCE.NORM. Figure 7-18 shows the calculations and formulae. Comparing this result with Figure 7-14, we note that the confidence limits are noticeably wider apart. Where the limits based on the parametric standard deviation were 732.0 and 758.0, the limits using the sample standard deviation are 730.4 and 759.6. The latter confidence interval is roughly 112% of the former confidence interval.

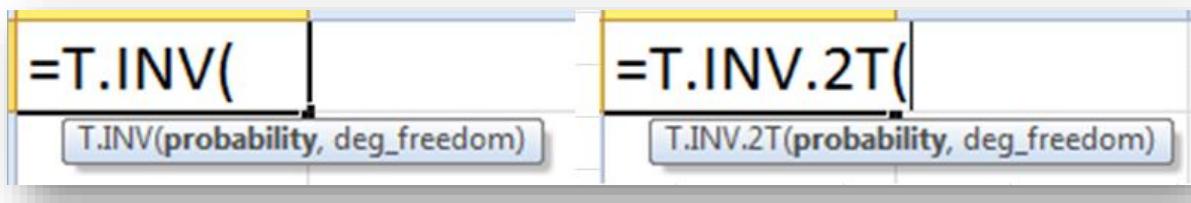
7.20 Beware the Definition of α in Inverse Probability Functions

There is another function that can be useful in many other applications, including computing confidence limits for other kinds of statistics than the mean. Many applications can use the =T.INV (or its older equivalent, =TINV) and the =T.INV.2T functions to compute critical values for $t_{\alpha/2}$.

Figure 7-19. Same word, different meanings.



However, a critical issue when computing critical values is that these functions – and other functions you will encounter in your explorations of EXCEL and of other statistical packages – have different, contradictory definitions of the probability parameter! Figure 7-19 shows the syntax of the EXCEL 2010 functions; notice that the parameter probability occurs in both.



- The =T.INV function yields a critical value using the left tail of the probability distribution, which means that if we enter =T.INV(.025, df), the function yields the *left-tail* critical value $t_{0.025} \downarrow$ (a negative number) which produces 0.025 on the *left* of the critical value and 0.975 to the right of the critical value. Because of the symmetry of the Student's-t distribution, that also means that the 0.025 of the distribution lies to the *right* of $|t_{0.025} \downarrow|$ and 0.975 of the distribution lies to the *right* of that value.
 - For example, we can calculate =T.INV(0.025, 100) = -1.98397. We write this as $t_{0.025 \downarrow [100]} = -1.98397$ or as $t_{0.975 \uparrow [100]} = +1.98397$
 - Thus the probability parameter in this function generates a critical value corresponding to a *one-tailed* probability for the *left-tail* critical value, a *negative* number.
- Now consider =T.INV.2T, which, as the .2T indicates, uses a *two-tailed* probability – and in addition, computes the *right-tail* critical value, a positive number.
 - For example, we can calculate =T.INV.2T(0.05, 100) = 1.98397. Exactly as above, we write this as $t_{0.025 \downarrow [100]} = 1.98397$. Notice that we have to describe it as cutting off a right tail with 0.025, not the 0.05 entered into the function!
 - So as you can see, the =T.INV.2T function automatically computes a critical value that defines the left tail and the right tail for the critical value as having *half the probability* entered in the probability parameter.

Remember this example: **you must verify whether a statistical function in any statistical package computes left-tail or right-tail critical values using one-tailed or two-tailed probabilities.** Don't get upset or curse the programmers for inconsistency: just check to make sure you are computing what you *need*, not what you *hope*.

If you want to check your understanding of a new function you haven't used before, you may be able to check your understanding using known results for known parameters to ensure that you are not mistaken in your understanding of what the new parameters mean for the new function.

7.21 Interval Estimate for Any Normally Distributed Statistic

Even more generally, we can extend the applicability of the t-distribution and its use in computing interval estimates of a parametric value to any normally distributed statistic.

Suppose we find a statistical research article that discusses the distribution of a new statistic, say, the “delta” coefficient (δ for parameters, d for samples). Imagine that this (made-up) statistic is important in evaluating the reliability of warp cores on starships. Extensive research confirms that the δ coefficient is indeed normally distributed. Starship engineers need to estimate the confidence limits for the parametric value δ given a sample’s d statistic because any value smaller than 3 or greater than 4 may lead to a faster-than-light engine implosion. Galaxyfleet insists on a chance of implosion of less than 5%.

The principle is that any *normally distributed* statistic (in this imaginary example, d) – whose standard error is s_d with v (for example, $v = n - 1$ degrees of freedom for a sample size of n) will fit the same pattern as what we have seen in computing confidence intervals for means using the Student’s-t distribution:

$$P\{d - |t_{\alpha/2[v]}| s_d \leq \delta \leq d + |t_{\alpha/2[v]}| s_d\} = 1 - \alpha$$

That is, interpreting this algebraic formulation,

- The probability P
- That we will be correct
- In asserting that the lower and upper computed $(1 - \alpha)$ confidence limits for δ
- With v degrees of freedom
- Include the parametric statistic δ
- Is $(1 - \alpha)$.

Example:

- An engineer finds that in a sample of 100 warp-core signatures, the sample d statistic is 3.48 and the standard error of d (s_d) with $v = 99$ degrees of freedom is 0.081.
- The two-tailed function $=T.INV.2T(.05,96)$ for $\alpha = 0.05$ (i.e., $\alpha/2$ in each tail) and $v = 100$ in EXCEL gives us the critical upper-tail value $t_{0.05[96]} = 1.984216952$ which in turn lets us compute the differential $t_{\alpha/2[v]} * s_d = 1.984216952 * 0.081 = 0.160721573$ or ~ 0.161 . So the limits of our interval estimate for δ are 3.48 ± 0.161 or 3.319 and 3.641.
- Thus we have a 95% chance of being correct in asserting that the interval 3.319 to 3.641 based on our sample values of d includes the parametric delta-coefficient δ .
- The confidence limits are within the range of acceptability, so the engineer concludes that the chances of a warp-core implosion today are less than 5%.

This example shows one of the ways that confidence limits can be used in quality control.

7.22 Population Proportion Based on Sample Proportion

In a study of the value of warning labels on pharmaceutical products, the BigPharma Association of the Greater Solar System looked at a sample of 2,000 sentients out of the total population of about 397,452,000 known to have been legally prescribed drugs with such warning labels and counted how many had bothered to read the labels. They found that the proportion p_{sample} of readers-of-labels was 22.5%. What was the 95% confidence interval for the parametric proportion $p_{\text{population}}$ of readers-of-labels?⁸²

Statisticians have shown that repeated measures of p_{sample} with sample size n are distributed as a Normal distribution with the mean $p_{\text{population}}$ (as expected under the Central Limit Theorem) and parametric variance

$$\sigma_p^2 = p(1-p)/n$$

provided that

- The population size N is infinite *or* that
- That the ratio of the sample size n to the population size N meets the condition

$$n/N \leq 0.05$$

- *And* that $np > 5$ and $n(1-p) > 5$

In other words, provided that the sample size n is less than 5% of the total sample size N , the approximation for the parametric variance of proportions works fine.

It follows that the parametric *standard error* of the sample proportion, σ_p is defined as

$$\sigma_p = \sqrt{\frac{p_{\text{sample}}(1-p_{\text{sample}})}{n}}$$

Once we know how to compute the standard error of the proportion and we know that the sample proportions are Normally distributed, we can compute the interval estimate for the population proportion using same principles described in §7.21 above.

Figure 7-20. Confidence limits for a proportion p .

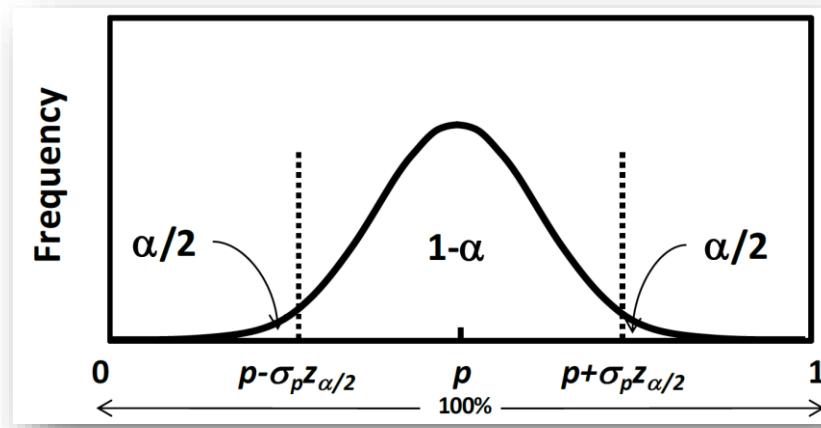


Figure 7-20 shows the meaning of the confidence limits graphically.

⁸² We don't use the Greek version of p (π) because it is too strongly associated with the ratio of the circumference to the radius of a circle in plane geometry. However, note that the capital version, Π , is frequently used to mean "product" in the same way that capital sigma, Σ , is conventionally used to mean "sum."

Figure 7-21 provides an example in EXCEL for computing *confidence limits for the proportion* of sentients who bothered to read the warning labels on pharmaceuticals in a study on planet Phlrn'thx in the Galaxyfleet year 2712. Column D in the spreadsheet displays the formulas used in Column B.⁸³

Figure 7-21. Demonstration of computing confidence limits for proportion.

	A	B	C	D
1	DATA INPUT		<i>Check</i>	<i>Formulas Displayed for Example Only</i>
2	total population N	397,452,000		
3	sample size n	2,000		
4	observed proportion p	0.225		
5	desired confidence level $1-\alpha$	0.95		
6				
7	PRELIMINARY CHECKS			
8	$n/N \leq 5?$	5.03E-06	OK	=+B3/B2 and in Check column =IF(B8<=5, "OK", "NO")
9	$np > 5?$	450	OK	=+B3*B4 and in Check column =IF(B9>=5, "OK", "NO")
10	$n(1-p) > 5?$	1,550	OK	=+B3*(1-B4) and in Check column =IF(B10>=5, "OK", "NO")
11				
12	CALCULATIONS			
13	α	0.05		=1-B5
14	$\alpha/2$	0.025		=B13/2
15	σ	0.009337425		=SQRT(B4*(1-B4)/B3)
16	critical z	1.9600		=ABS(NORM.S.INV(0.025))
17	half the confidence interval	0.0183		=(\$B\$15*\$B\$16)
18	lower 95% confidence limit	0.2067		=\$B\$4-\$B\$17
19	upper 95% confidence limit	0.2433		=\$B\$4+B17
20				
21	POST-CALCULATION CHECKS			
22	proportion below lower limit	0.025	OK	=NORM.DIST(+\$B\$17,\$B\$4,\$B\$15,1) and in Check column =IF(B21=\$B\$14, "OK", "NO")
23	proportion above upper limit	0.025	OK	=1-NORM.DIST(+\$B\$18,\$B\$4,\$B\$15,1) and in Check column =IF(B22=\$B\$14, "OK", "NO")
24				

The **CHECKS** sections verify that the results make sense. It's always worth checking your formulas the first time you use them by using information you know. In this case, the formulas in the **PRELIMINARY CHECKS** compute the guidelines and verify that they are within specification. **POST-CALCULATION CHECKS** go backward from the computed confidence limits to verify that the proportion of the curve below the lower confidence limit and above the upper confidence limit match $\alpha/2$. As it should be, the proportions are equal to $\alpha/2$; they also add up to 1.000, as they must.

The method discussed above works as an approximation that is acceptable for proportions that are not very close to zero or to one – something verified by the **$np > 5$** and **$n(1 - p) > 5$** conditions in the **PRELIMINARY CHECKS** section. Later in the course, you will learn about other methods of determining confidence limits for proportions that don't fit these assumptions.

⁸³ If you ever need to display the formulas and the results in the same sheet, you use F2 to enter EDIT mode, copy the formula, type an apostrophe in the target cell and paste the formula into place right after the apostrophe. It will then be a string rather than an active formula. For example, Cell D3 actually contains '=+B2/B1' and but it does not display the leading apostrophe.

7.23 Conditional Formatting

In Figure 7-21, you may have noticed the green boxes with OK in them next to the checks. These boxes are formatted with **CONDITIONAL FORMATTING** in EXCEL 2010. Figure 7-22 shows the drop-down menu for **CONDITIONAL FORMATTING**.

Conditional formatting determines the appearance of a cell according to a wide range of possible conditions or rules. There are a great many options in **CONDITIONAL FORMATTING**, but Figure 7-23 demonstrates how a simple formula can warn a user visually that something is wrong.

In this example, the check-result cells (C8 through C10, C21 and C22) are initially tinted light green by default with dark green letters. If the contents (defined by an =IF statement) are “NO” then the box turns light red and the letters are deep red.

Defining appropriate conditional formatting, especially for a spreadsheet that you plan to use often or that you are putting into production for other people to use, can instantly warn a user of an error. Although the colors are helpful, even a color-blind user can see an appropriate message (e.g., “NO” or “BAD” or “ERROR”) to signal something wrong.

The EXCEL 2010 **HELP** function has a number of useful articles about *conditional formatting* that you can access by entering that term in the search box.

Figure 7-22. Accessing Conditional Formatting for an existing rule.

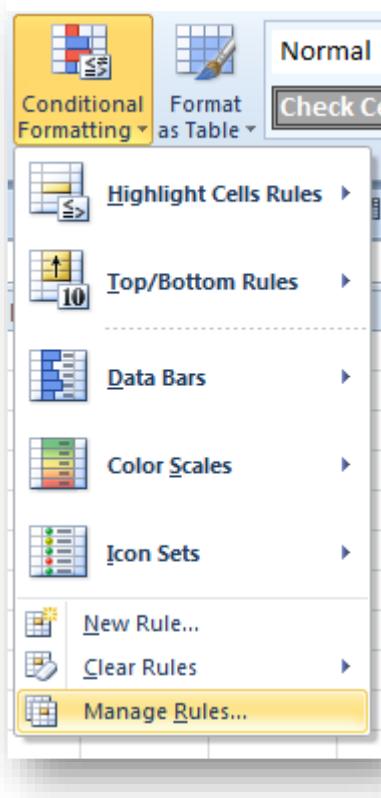
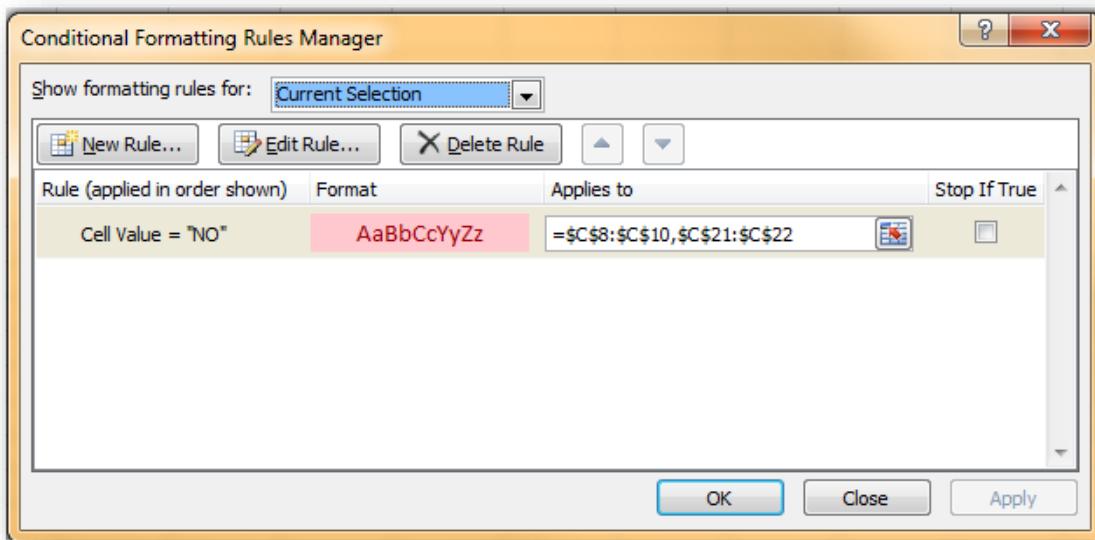


Figure 7-23. Definition of rule for Check cells in example.



7.24 Confidence Limits for Population Variance and Population Standard Deviation Based on Sample Variability

Erewham Natural Foods Corporation are sampling grandiloquent beetle carapaces to make a medicinal grandiloquent beetle carapace extract (GBCE™) highly popular among outworld populations such as the Drazeeli and the Q'ornoppiads. The company is deeply concerned about the quality control in its Io plant because deviation from its contractual obligation can result in executive decapitation (on Drazeel) and slow conversion into compost (on Q'ornopia). For the Erewham plant to pass ISO (Interplanetary Standards Organization) 9000 standards, it must monitor the standard deviation of its 1000 gram bottles and start investigating the production line when the standard deviation of the production exceeds a parametric standard deviation of 2 grams.

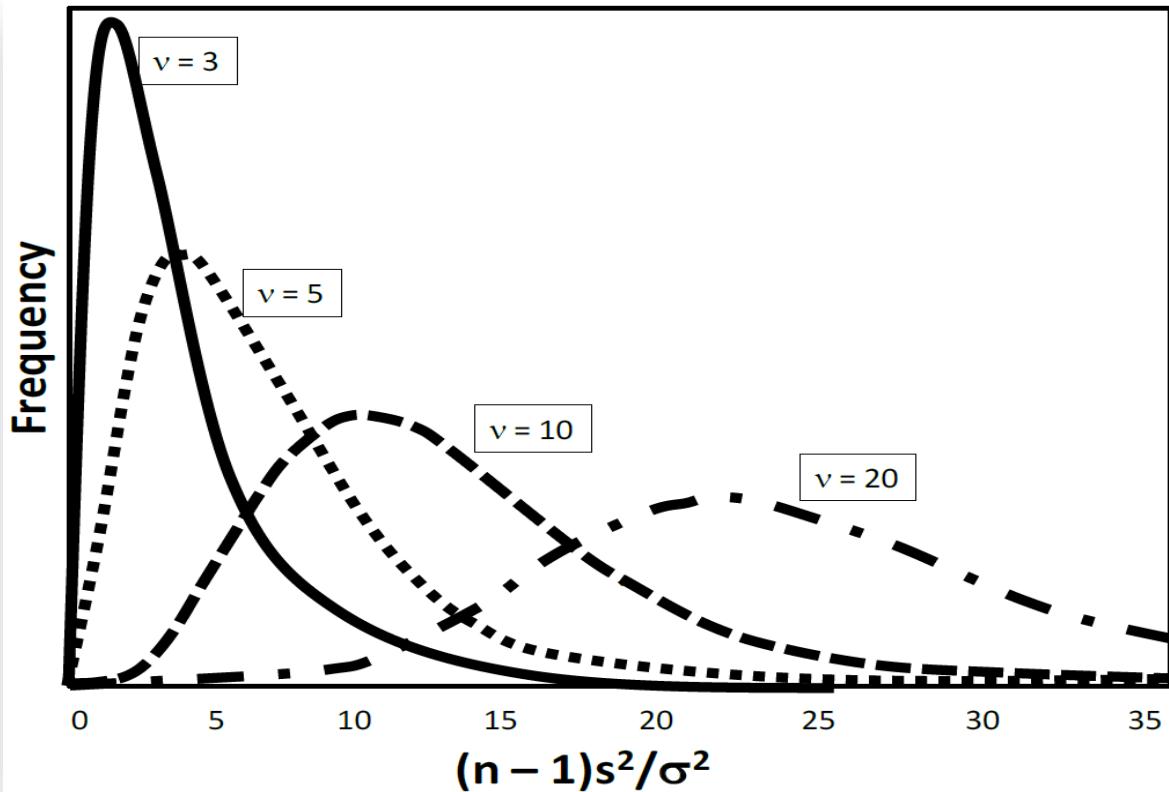
- For extra care in maintaining its corporate reputation, every day the plant managers compute 99% confidence limits for the standard deviation
- Using a random sample of 25 bottles of GBCE™ selected from output using random numbers.
- On a particular day in October 2281, the sample comes back with a standard deviation of 0.8 gm.
- What are the 99% confidence limits for the standard deviation that day?

Variances and standard deviations are *not* normally distributed. Instead, the quantity

$$\chi = (n - 1)s^2/\sigma^2 = \nu s^2/\sigma^2$$

(where n is the sample size, s^2 is the sample variance, and σ^2 is the parametric variance) is distributed according to a theoretical distribution called the *chi-square* (χ^2) with $\nu = (n - 1)$ degrees of freedom. Several of these distributions are shown in Figure 7-24 with different degrees of freedom.

Figure 7-24. Chi-square distributions with different degrees of freedom.

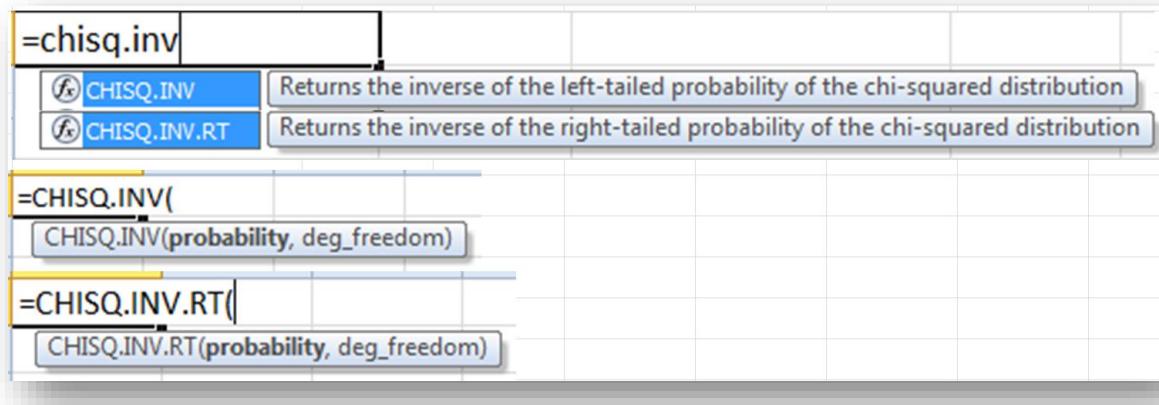


By convention, we use the notation $\chi^2_{\alpha[v]}$ to designate the critical value of the χ^2 distribution with v degrees of freedom for which the probability of sampling a chi-square variable x greater than critical value is α ; i.e., by definition

$$P\{x \geq \chi^2_{\alpha[v]}\} = \alpha$$

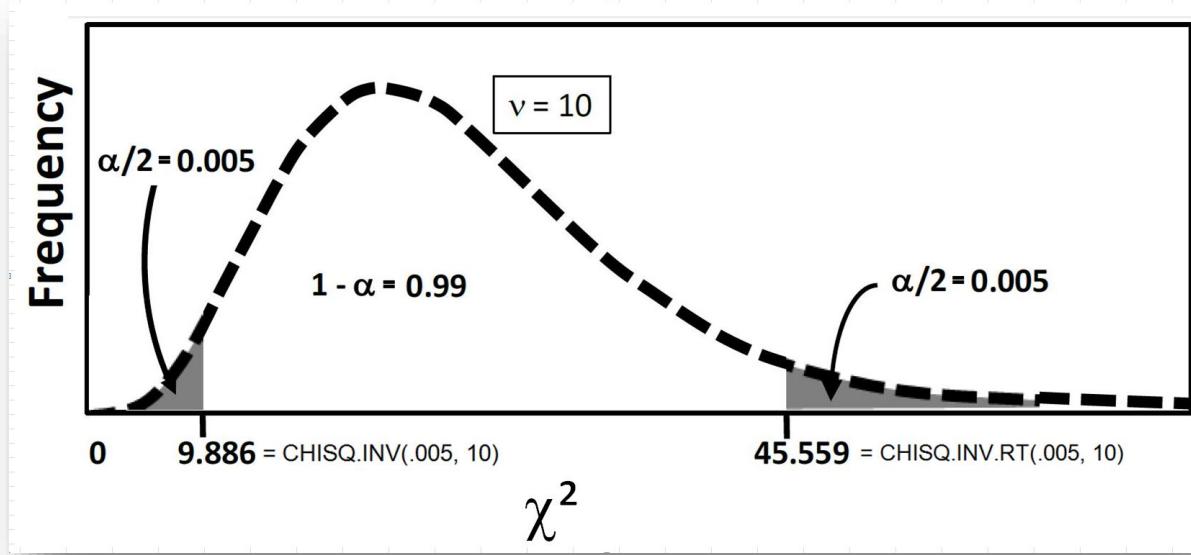
There are two chi-square inverse functions – one for each tail – in EXCEL 2010 that can provide the critical values needed for computation of confidence limits to a variance. Figure 7-25 shows a compound screenshot of these variables in EXCEL:

Figure 7-25. Excel 2010 popups for chi-square inverse functions.



- The EXCEL function `=CHISQ.INV(x, deg_freedom)` gives the *left-tail* critical value $x = \chi^2_{\alpha[v]}$ of the χ^2 distribution with $df = v$. For example, the critical value $\chi^2_{.005[10]} = \text{CHISQ.INV}(.005, 10) = 9.886$
- The EXCEL function `=CHISQ.INV.RT(x, deg_freedom)` generates the right-tail critical value corresponding to the value x ; thus $= \text{CHISQ.INV.RT}(.005, 10) = 45.559$
- Thus the $(1 - \alpha) = 0.99$ confidence limits imply $\alpha/2 = 0.005$ in each of the left and right tails of the distribution, and the $(1 - \alpha) = 0.99$ in the area in the middle, as shown in Figure 7-26.

Figure 7-26. Critical values of the chi-square distribution for computation of confidence limits.



To compute confidence limits to the parametric variance based on a sample variance, we need to expand our earlier definition of the critical value.⁸⁴ We know that for the $(1 - \alpha)$ confidence limits, we can start with

$$P\{ \text{CHISQ.INV}(\alpha/2, v) \leq x \leq \text{CHISQ.INV.RT}(\alpha/2, v) \} = 1 - \alpha$$

Substituting the meaning of x ,

$$P\{ \text{CHISQ.INV}(\alpha/2, v) \leq vs^2/\sigma^2 \leq \text{CHISQ.INV.RT}(\alpha/2, v) \} = 1 - \alpha$$

$$P\{ \text{CHISQ.INV}(\alpha/2, v)/vs^2 \leq 1/\sigma^2 \leq \text{CHISQ.INV.RT}(\alpha/2, v)/vs^2 \} = 1 - \alpha$$

Therefore⁸⁵

$$P\{ vs^2/\text{CHISQ.INV}(\alpha/2, v) \geq \sigma^2 \geq vs^2/\text{CHISQ.INV.RT}(\alpha/2, v) \} = 1 - \alpha$$

Or, putting the expression back in the normal order,

$$P\{ vs^2/\text{CHISQ.INV.RT}(\alpha/2, v) \leq \sigma^2 \leq vs^2/\text{CHISQ.INV}(\alpha/2, v) \} = 1 - \alpha$$

This is the general form for the $(1 - \alpha)$ confidence limits to the variance given a sample variance, with the lower limit (L_1) using the *right-tail* critical value (because it's larger and so the quotient is smaller) and the upper limit (L_2) using the *left-tail* critical value (because it's smaller and so the quotient is bigger).

Thus the confidence limits for the parametric variance based on the sample variance are computed as shown below using the functions in EXCEL 2010:

$$L_1 = vs^2/\text{CHISQ.INV.RT}(\alpha/2, v)$$

$$L_2 = vs^2/\text{CHISQ.INV}(\alpha/2, v)$$

The confidence limits for the *standard deviation* are the *square roots of the limits of the variance*.⁸⁶

INSTANT TEST P 7-31

A security engineer is trying to establish confidence limits for the variance and of the standard deviation for the number of input-output (I/O) operations per minute on a particular network storage unit under normal load.

The sample size is 10,000 and the observed sample standard deviation is 26.8.

Demonstrate that the lower and upper 95% confidence limits for the variance are 699 and 739; the lower and upper 95% confidence limits for the standard deviation are 26.4 and 27.2. At the 95% level of confidence, why is an observed variance of 600 unlikely?

⁸⁴ You are not expected to be able to derive these formulas from memory. They are presented to help all students understand the logic behind the computations and to support those students with an interest in the mathematical underpinnings of statistics. Courses in statistics offered in mathematics departments in universities usually include derivations of this kind for all computational formulas.

⁸⁵ Note the reversal of direction: if $2 < x < 3$ then $1/x > 1/3$.

⁸⁶ Neither a variance nor a standard deviation may be negative.

To illustrate these computations, we can return to the question of quality control of GBCE™ at the Erewham Company production line introduced at the start of this section. The Erewham data produce the following values using some simple EXCEL calculations shown in Figure 7-27.

Figure 7-27. Excel 2010 calculations of confidence limits for the variance and the standard deviation based on sample data.

A	B	C	D	E	F
1	Data & Calculations		Symbol	Value	Formula
2	Sample size	n		25	
3	Degrees of freedom	v		24	
4	Standard deviation of sample	s		0.8	
5	Variance of sample	s^2		0.64	=E4^2
6	Numerator for limits	$(n-1)s^2$		15.36	=E3*E5
7	Confidence level	$(1 - \alpha)$		0.990	
8	Left-tail probability	$\alpha/2$		0.005	=(1-E7)/2
9	Right-tail probability	$1 - \alpha/2$		0.995	=1-E8
10	Chi-square for lower limit			9.886	=CHISQ.INV(E8,E3)
11	Chi-square for upper limit			45.559	=CHISQ.INV.RT(E8,E3)
12	Confidence limits				
13	For Variance				
14		Lower		0.337	=E6/E11
15		Upper		1.554	=E6/E10
16	For standard deviation				
17		Lower		0.581	=+SQRT(E14)
18		Upper		1.246	=+SQRT(E15)
19					
20	CHECKING:				
21	Left tail for left-tail critical value			0.005	=CHISQ.DIST(E10,E3,1)
22	Right tail for right-tail critical value			0.005	=CHISQ.DIST.RT(E11,E3)

The 99% confidence limits for the standard deviation are 0.581 to 1.246 grams⁸⁷ but the maximum acceptable standard deviation with 99% confidence is much higher, at 2 grams. There's no reason to worry about decapitation or conversion into compost today!

Quality-control (QC) charts typically mark the upper and lower limits to the statistics being monitored and graph the periodic measures; in this case, $L_2(\sigma)$ of 1.246 grams would be below the maximum allowable standard deviation of 2 grams. Production goes on and the Solar System will continue to benefit from grandiloquent beetle carapace extract (GBCE™).

⁸⁷ A quick note about asymmetric confidence limits: the midpoint of 1.246 and 0.581 is 0.932, which is *not* the observed standard deviation of 0.8; the same asymmetry affects the variance limits, where the midpoint of the limits is 0.946, also not the observed variance of 0.64. The reason is that the underlying probability distribution (χ^2) is asymmetrical. Don't expect all confidence limits to be symmetrical around the sample value!

8 Hypothesis Testing

8.1 Introduction

The Ionian GBCETM confidence limits for the standard deviation in §7.24 came very close to asking whether the sample standard deviation could have come from a population whose parametric standard deviation was greater than 2. Hypothesis testing in statistics is usually in the form of the question, “Could the results we observe in our sample have occurred by chance variation in sampling alone if one or more parameters had specific values?” The hypothesis that defines the baseline against which the sample is evaluated is called the *null hypothesis* (H_0 or H_0 in most textbooks) and the particular hypothesis that will be accepted if we reject the null hypothesis is called the alternative hypothesis (most commonly H_1 , H_1 or H_a).⁸⁸

Typically, the question of *interest* will be represented by the alternative hypothesis. As illustrated in the following examples, note how consistently what is interesting to the analyst is the alternative hypothesis in the following examples of some questions we might encounter and the corresponding statistical hypotheses that might be framed:

- An accountant doing an audit is becoming suspicious of the figures shown in the books of a big company called Unron; she extracts the data from several hundred transactions and wants to know if the frequencies of the ten digits (0, 1,...9) in the last portions of the entries are equal (radical deviation from equality would suggest that the numbers were fraudulently invented, since people aren't very good at making up numbers that fit the uniform probability distribution).
 - H_0 : the frequencies *are* all 0.1;
 - H_1 : the frequencies *are not* all 0.1.
- A stock broker has become interested in the performance of the shares for Macrohard Corporation; he wants to know if the data for the last three years support the view that the growth rate is at least 6% per year. If it is, he will recommend to a client interested in long-term investments that the investment fits the client's profile.
 - H_0 : the regression coefficient is *less than* 6% per year.
 - H_1 : the regression coefficient of stock price versus year is *6% per year or more*;
- A network management team needs to know if a new heuristic compression algorithm is working *better* than the old Lempel-Zev-Welch (LZW) algorithms that have been in use for decades. They measure the compressed file sizes for a wide range of files using both types of algorithms and compare the results.
 - H_0 : there is *no difference* between the compression ratios of the LZW compression methods and the new heuristic algorithms OR the heuristic algorithms are *not as good as* the LZW methods.
 - H_1 : the heuristic algorithms are *better* than the LZW algorithms.

⁸⁸ This text uses H_0 and H_1 to avoid constantly having to typeset subscripts, especially in Excel, where subscripts are a real pain to create. In exercises, students are exposed to both H_1 and H_a for practice.

- A manufacturer of trans-temporal framrigers is investigating the effects of besnofring modulation on the accuracy of the time-tuning mechanism. The investigators apply seven different levels of besnofring modulation to the framrigers while transporting samples to ten different time locations each. They analyze the 70 measurements to see if there are any effects of the besnofring modulation.
 - H0: there are *no effects* of differences in besnofring modulation level on the accuracy of time-tuning.
 - H1: there *are* differences in accuracy of time-tuning among the groups exposed to different levels of besnofring modulation.
- A marketing firm has three different variations on an advertising campaign. They are trying them out in six different regions of the target market by counting the number of answers to questions using a Likert scale (1: strongly disagree, 2: disagree... 5: strongly agree).
 - Are there differences among the ad versions in the way people respond in the six regions overall?
 - H0: there are *no* main (overall) effects of ad versions on responses;
 - H1: there *are* main effects of ad versions on responses.
 - Are there differences among the regions of the market in the way people respond?
 - H0: there are *no* main effects of region on responses;
 - H1: there *are* main effects of region on responses.
 - Are there any regional variations in the way people respond to the different campaigns
 - H0: there are *no interactions* between the ad variations and the regions of the market in the way people respond;
 - H1: there *are* interactions between the ad variations and the regions of the market.
- An investor is looking at two different manufacturers of trans-temporal framrigers as potential investments. One of the steps in due diligence is to examine the reliability of quality control of the two factories' production lines by comparing the variances of the products.
 - H0: there is no difference in the variances of the two production lines;
 - H1: there is a difference in the variances of the two production lines.
- A system manager needs to know if a particular department's growth rate in disk space utilization is really faster than all the other departments' growth rates.
 - H0: the regression coefficient of disk space over time for the suspect department is *not different* from the other departments' regression coefficients or it is *slower*.
 - H1: the regression coefficient of disk space over time for the suspect department is *greater* than the other departments' regression coefficients.

In general, the null hypothesis, H0, is the one that posits no effect, no difference, no relationship, or a default state. H1, the alternative hypothesis, is the one that posits an effect, a difference, a relationship, or deviation from a ho-hum uninteresting state. There is nothing absolute or inevitable about the framing of H0 and H1: the decisions depend on the interests of the investigator.

8.2 Are the Variances of these Two Samples the Same?

Although most statistics textbooks start with comparisons of the means, it will be very useful for you to know about testing variances – the test is used in analysis of variance (ANOVA), which is a central technique in applied statistics. Here we begin the study of hypothesis testing by looking at variances.

Statisticians have determined that when we repeatedly take two samples from the same population and compare their variances, the ratio, called an F-statistic, follows a distribution called the F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. The degrees-of-freedom parameters apply to the numerator and the denominator of the ratio.

For the ratio of two sample variances, s^2_1 and s^2_2 computed the usual way from samples of size n_1 and n_2 , we compute

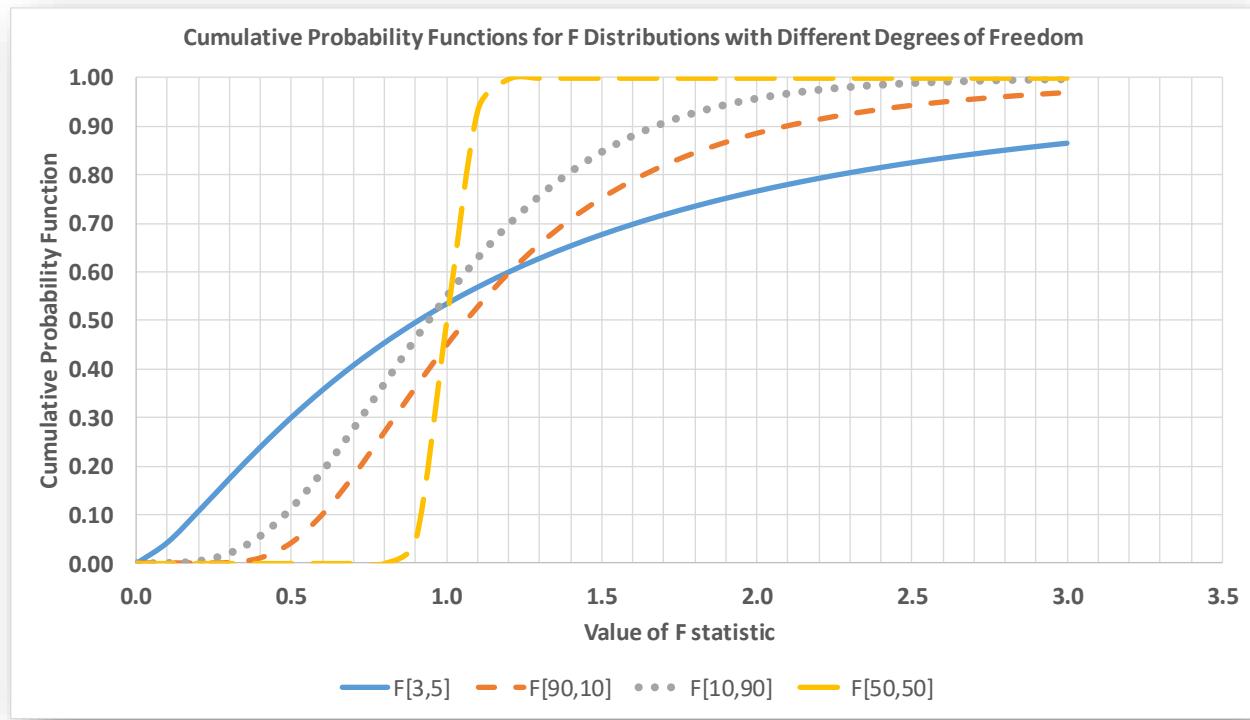
$$F_{v_1 v_2} = \frac{s^2_1}{s^2_2}$$

This *F-ratio* follows the F distribution with v_1 and v_2 degrees of freedom.

Each combination of degrees of freedom has a different shape, as shown in Figure 8-1, which graphs cumulative probability functions for four combinations of n_1 and n_2 degrees of freedom.

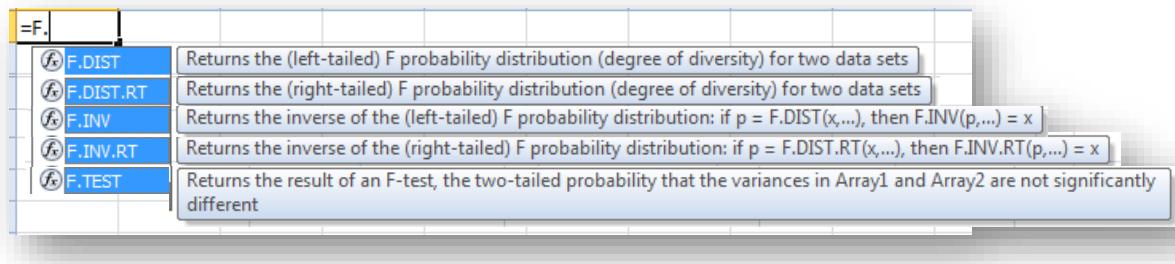
The F-test for the equality of two sample variances (or of any other quantity that follows the same rules as variances) consists of computing the ratio the sample variances and determining how often such a large value of F or larger would occur by chance alone (i.e., through random sampling) if both samples actually came from the same population or from two populations with identical parametric variances.

Figure 8-1. Cumulative probability functions for various F distributions.



By convention, we create the F-ratio by dividing the larger sample variance by the smaller sample variance so that F will be 1 or larger, allowing us to use F.DIST.RT (=FDIST in earlier versions) to compute the right-hand tail of the F-distribution for the appropriate pair of degrees of freedom. The EXCEL function =F.DIST.RT(F, n1, n2) provides the probability that the value F (and any values larger than the one observed) of the F-ratio would occur by chance if the samples came from the same population or if they came from two populations with identical parametric variances. The F-distribution functions are shown in the composite image in Figure 8-2.

Figure 8-2. Excel 2010 F-distribution functions.



For example, =F.DIST.RT(2.5,10,100) = 0.010; thus the probability of obtaining an F-ratio of 2.5 or larger by chance alone when the degrees of freedom of the numerator and denominator are 10 and 100, respectively, is 0.010.

If for some reason you are given an F-ratio that is smaller than 1, you can use the =F.DIST function to calculate the left-tail probability. For example, if the F-ratio were inverted, giving =F.DIST(0.4,100,10,1),⁸⁹ the result would be 0.010, as it should be.

There are three possible sets of hypotheses that you might want to test with an F-test:

- $H_0: \sigma^2_1 = \sigma^2_2$ and $H_1: \sigma^2_1 \neq \sigma^2_2$
- $H_0: \sigma^2_1 \leq \sigma^2_2$ and $H_1: \sigma^2_1 > \sigma^2_2$
- $H_0: \sigma^2_1 \geq \sigma^2_2$ and $H_1: \sigma^2_1 < \sigma^2_2$

Suppose we are interested in the political commitment of two different groups of voters in the Martian general elections, the Mars First Coalition secessionists and the Earth Forever Patriots irredentists. One measure of their commitment is the variability of their responses to ideological statements reflecting each movement's core values. A University of Schiaparelli political science group develops test instruments and applies them to 1,228 Mars Firsters (secessionists) and to 1,175 Earth Forever (irredentists) members. Their results show that the variance of the overall scores are 38.2 for the secessionists and 35.0 for the irredentists. Are there grounds for believing that the two groups have a difference between their parametric variances?

Calculate the F-ratio that is greater than 1:

$$F_{[1227,1174]} = 38.2/35.0 = 1.091429$$

Then

$$P\{ F_{[1227,1174]} \geq 1.091429 \mid \sigma^2_1 = \sigma^2_2 \} = F.DIST.RT(1.091429, 1227, 1174) = 0.06504$$

We can read the line above as "The probability of obtaining an F-ratio of 1.091429 or larger by chance alone using these sample sizes if the two population variances were equal is 0.0650."⁹⁰

⁸⁹ For unknown reasons, Microsoft decided to include the *cumulative* parameter (1=cumulative, 0=density) in the =F.DIST function but not in the =F.DIST.RT function.

⁹⁰ $P\{ a \mid b \}$ is called a *conditional probability* and is interpreted as "The probability of observing *a* given that *b* is true."

8.3 Levels of Statistical Significance and Type I Error: Rejecting the Null Hypothesis When it is Actually True

At this point, a reasonable question is “So what if the probability of the F-ratio is 0.0650? What does that mean to us? Are the parametric variances different or aren’t they?”

A conventional answer points to four agreed-upon levels of statistical significance in hypothesis testing. We conventionally refer to the probability of obtaining the observed results by chance alone if the null hypothesis is true as $P\{H_0\}$ and even more often simply as p . We also call p the probability of Type I Error.

Type I Error is *rejecting the null hypothesis when it is actually true*.

We agree by convention that

- If $p > 0.05$
 - We accept H_0 ,
 - Reject H_1 , and
 - Say that the results are *not statistically significant* and
 - Follow the $P\{H_0\}$ (or p) with the letters *ns* to show non-significance; e.g., “The probability of encountering such a large statistic or larger if the parametric variances were the same is 0.0650, which is not statistically significant at the 0.05 level of significance. We therefore accept the null hypothesis of equal variances for these two samples.”
- If $0.01 < p \leq 0.05$
 - We reject H_0 ,
 - Accept H_1 , and
 - Say that the results are statistically significant or statistically significant at the 0.05 level.
 - The $P\{H_0\}$ (or p) is typically marked with one asterisk (*) to show this level of significance; e.g., $P\{H_0\} = 0.0472^*$
 - The most explicit formulation of the results is something like this: “The test statistic of 34.98 has a probability of 0.0472* of occurring by chance alone even if the samples had the same parametric value of this statistic. The results are statistically significant at the 0.05 level of significance and we reject the null hypothesis of equality of the parametric statistics.”
 - In practice, we would be more likely to write briefly, “The parametric statistics for these samples are significantly different at the 0.05 level of significance ($p = 0.0472^*$).”
- If $0.001 < p \leq 0.01$
 - we reject H_0 ,
 - accept H_1 , and
 - Say that the results are statistically highly significant or statistically significant at the 0.01 level.
 - The $P\{H_0\}$ is typically marked with two asterisks (**) to show this level of significance; e.g., $P\{H_0\} = 0.00472^{**}$.
 - The verbose formulation could be, ““The test statistic of 87.02 has a probability of 0.00472** of occurring by chance alone even if the samples had the same parametric value of this statistic. The results are statistically significant at the 0.01 level of significance and we reject the null hypothesis of equality of the parametric statistics.”
 - In practice, we would probably write something like, “The parametric statistics for these samples are highly significantly different at the 0.01 level of significance ($p = 0.00472^{**}$).”

- If $p \leq 0.001$
 - We reject H_0 ,
 - Accept H_1 , and
 - Say that the results are statistically extremely significant or statistically significant at the 0.001 level.
 - The $P\{H_0\}$ is typically marked with three asterisks (***) to show this level of significance; e.g., $P\{H_0\} = 0.000472***$.
 - The verbose formulation could be, “The test statistic of 109.53 has a probability of 0.000472** of occurring by chance alone even if the samples had the same parametric value of this statistic. The results are statistically significant at the 0.001 level of significance and we reject the null hypothesis of equality of the parametric statistics.”
 - In practice, we would probably write something like, “The parametric statistics for these samples are extremely significantly different at the 0.001 level of significance ($p = 0.000472***$).”

So going back to the study of Martian colony politics, are the two political movements similar or different in their variability on their ideological position?

The way a statistician would normally respond to the test results would be to say that the results of the study were not statistically significant, since the probability of rejecting the equality of variances even though they might be the same was about 6%, above the normal cutoff point of 5% for statistical significance. However, taking into account the size of the samples (respectable) and the closeness of the p -value to the limit for significance, a statistician would also add that the results are *highly suggestive* even though not statistically significant and that the problem could bear further study.

The issue of whether experiments or studies are repeated if the initial results don't fit preconceptions or preferences is a, you should pardon the term, significant problem in statistics. Even without a theoretical analysis, it must be clear that repeating studies until one gets the result that's desired – and ignoring the contrary results of the earlier studies – is surely going to bias the results towards the preconceived goal. A reasonable approach must repeat experiments if they are close to a minimum regardless of whether they are on one side or another. Thus a good experimental protocol might include “The experiment will be repeated if the p -value is from **0.02** through 0.08.” What *won't* work is “The experiment will be repeated if the p -value is from **0.05** through 0.08.”

Here's another example of a test of variances, this time drawn from computer science.

A musician working with machine intelligence has created a computer program that analyzes samples of music from specific composers – sometimes thousands of compositions – and creates complex algorithms reflecting the composers' patterns, including the degree of spontaneity and originality, in their music. Using these algorithms, the program interacts with its operator to construct new music that can be in the style of the original composer or that can be wholly original. The musician is studying listeners' response to the music and is interested in knowing if the range of responses is different when they listen to the original composers' music compared with the range of responses when they listen to the synthesized music. Thus

$$H_0: \sigma^2_1 = \sigma^2_2 \quad \text{and} \quad H_1: \sigma^2_1 \neq \sigma^2_2$$

The researcher collects data on 40 people's responses and finds that the F-ratio of the variance of the responses to the original music compared to the variance of the responses to the synthesized music is 1.83 with 39 and 39 degrees of freedom and $P\{H_0\} = 0.0314*$. We can say “There is a statistically significant greater variance of the responses to the original music compared with the variance of the responses to the synthesized music ($p = 0.0314*$).”

As we have seen, we often encounter assertions of the form, “If the null hypothesis is true, the chance of observing this statistical result or one even more extreme is 0.042*, which is statistically significant at the 0.05 level of significance.”

But what if we are unlucky? What is the probability of observing the results we obtained if the null hypothesis is actually true but we have stumbled into the improbable case? We simply assert that if the probability that the null hypothesis is p , then the risk of *rejecting* the null hypothesis when it is *true* is simply p . So in the example above, the probability p of encountering the observed results if H_0 is true is 0.042 and the probability that we will be wrong in accepting the null hypothesis is exactly that: 0.042.

8.4 Type II Error: Accepting the Null Hypothesis when it is Actually False.

In general, if we choose the α level of significance, we have to understand that we will be *wrong* on average in α of the statistical decisions we make that accept the null hypothesis. Thus if we choose $\alpha = 0.01$, there is a 1% chance that we will encounter the deviant results that will trick us into rejecting the null hypothesis when it is true.

It follows that if we increase the severity of our test criterion – e.g., if we use $\alpha = 0.001$ as our limit for rejecting the null hypothesis, there is an increasing risk that we will *accept* the null hypothesis when it is actually false. We call this kind of error *Type II Error*.

Figure 8-3. Type I and Type II Errors.

Decision Under Uncertainty	H_0 TRUE	H_0 FALSE
Accept H_0	CORRECT	TYPE II ERROR
Reject H_0	TYPE I ERROR	CORRECT

The calculation of the probability of Type II Error (sometimes symbolized β) is complicated because it's impossible to calculate theoretical probabilities of the observed values without specifying an explicit value for the null hypothesis. If H_0 assumes equality of means, there is an infinite number of versions of H_0 in which the means differ.

Statisticians have developed methods for estimating the probability of Type II Error, but the topic is suitable for a more advanced course than this introduction.

It is enough for the time being for students to be aware that there is always a possibility of being wrong in our *decisions under uncertainty*.

8.5 Testing a Sample Variance Against a Parametric Value

It is possible that you may have to test the variance of a specific sample, s^2 , against a known or assumed parametric value, σ^2 . You may recall the example in §7.24, where we developed a confidence interval for the parametric variance based on a sample statistic: the situation involved production of 100 gram bottles of a beetle-carapace extract at an Erewash plant, where the rules were to start investigating the production line if the standard deviation of the sample exceeded a parametric limit of 2 grams. One approach was to compute confidence limits; now you can see that another approach is to test the hypothesis that the parametric standard deviation of the population from which a sample has been extracted is equal to 2 grams or less. If so, no problem; if not, problem! Exceeding the maximum allowable variance could lead to throwing the entire production into the compost bin.

We have to convert our standard deviations into variances to be able to calculate the test value. So $s^2 = 0.8^2 = 0.64$ and $\sigma^2 = 2^2 = 4$ and the null and alternative hypotheses in this case would be

$$H_0: \sigma^2 \leq 4 \quad \text{and} \quad H_1: \sigma^2 > 4.$$

In §7.24, you were told that the quantity

$$x = (n - 1)s^2/\sigma^2$$

follows a chi-square distribution with $df = (n - 1)$ degrees of freedom. This quantity thus becomes the basis of a hypothesis test for sample variances compared to parametric variances.

We compute the quantity x and compare it to the $\chi^2_{[n-1]}$ distribution to determine the likelihood that the x value (or a value even more unexpected – i.e., larger) could have occurred by random sampling alone if the sample really came from a population with the defined parametric variance $\sigma^2 = 4$.

Let's continue with our beetle carapace example and look at the results of the sample discussed in §7.24. The quality-assurance test took 25 bottles of the hugely popular *Grandiloquent Beetle Carapace Extract* (GBCE™) from Vega II on a particular day in 2281 and found the sample standard deviation to be 0.8 gm. Summarizing the situation, we have

$$n = 25 \quad s = 0.8 \text{ and so } s^2 = 0.64 \quad \sigma = 2 \quad \text{and so } \sigma^2 = 4$$

Therefore

$$x = (n - 1)s^2 / \sigma^2 = 24 * 0.64 / 4 = 3.840$$

and this quantity is distributed as the $\chi^2_{[24]}$ distribution if the data are the result of random sampling from a population (production line) with a parametric variance of 4 (parametric standard deviation of 2).

Using the CHISQ.DIST.RT function in EXCEL 2010,

$$P\{H_0\} = \text{CHISQ.DIST.RT}(3.94, 24) = 0.9999988302 \approx 1$$

in other words, it is *almost certain* that this sample could have been drawn from a population (the production for that day) with a parametric variance of 4 (i.e., a parametric standard deviation of 2). Therefore, we accept the null hypothesis: $\sigma^2 \leq 4$. Production of beetle carapace extract can continue without risk of conversion of the product to compost (what a relief)!

Just for practice, what if the sample standard deviation s had actually been 3.0 instead of 0.8? Then $s^2 = 9$ and

$$x = (n - 1)s^2 / \sigma^2 = 24 * 9 / 4 = 54 \text{ and}$$

$$P\{H_0\} = \text{CHISQ.DIST.RT}(54, 24) = 0.000426^{***}$$

Technically, we would write that “The probability that we could obtain a sample with standard deviation of 3 or more if the parametric standard deviation were 2 or less is only 0.0004*** which is extremely statistically significant at the 0.001 level of significance.”

8.6 Are the Means of These Two Populations the Same?

In every field of investigation, people need to know if the results of their measurements in two different samples could have come from the same population . Are these samples from the same population and therefore different due to random sampling alone? Or are they from two different populations whose parametric means differ? One of the most powerful tools in applied statistics is the analysis of variance, or ANOVA, which builds on what we have learned about testing the equality of sample variances. Another widely-used tool is the t-test of the equality of two statistics that follow a Normal distribution.

8.7 ANOVA for Comparing Means of Two Samples

Imagine that we have two samples of size 30 that we know are actually drawn from the *same population of measurements* of a production line of 10" radius tires for very small environmentally-clean cars powered by pet rabbits. We need to find out if the means of the two samples reflect differences in the parametric means of the populations sampled or if we can accept that the means as equal (H_0).

The mean radius and variance for each sample are as follows (the unusual precision is provided so readers can do their own computations):

Figure 8-4. Mean and variance of two samples.

Samples	Mean Radius	Variance of radius
Sample 1	9.992921717	0.08067098
Sample 2	10.03033012	0.08771458

Before we can use ANOVA to test the significance of the observed difference between the means, we should satisfy ourselves that one of the key assumptions of ANOVA is satisfied: the parametric variances of the samples must not differ. Using our usual notation, we must test

$$H_0: \sigma^2_1 = \sigma^2_2 \quad \text{and} \quad H_1: \sigma^2_1 \neq \sigma^2_2$$

We know, as godlike beings, that there are really no differences in the parametric statistics of these two samples. They are just samples from the same population. The statisticians working with the data, however, don't know this: they have to manage with probability calculations.

Noticing that the variance for the second sample is bigger than that of the first sample, we compute the F-test for equality of the parametric variances of the populations from which these samples are drawn as

$$F_{[29,29]} = s^2_2 / s^2_1 \text{ to generate a value } > 1 \text{ for the test.}$$

$$F_{[29,29]} = 0.08771458 / 0.08067098 = 1.087$$

How likely is it that we could observe an F-ratio that large or larger if it were true that $H_0: \sigma^2_1 = \sigma^2_2$?

We compute

$$= F.DIST.RT(1.087, 29, 29) = 0.412ns$$

and thus conclude that there is no reason to reject $H_0: \sigma^2_1 = \sigma^2_2$. We may proceed with the test of the means.

Now suppose we combine the two samples into one big sample with $n = 60$ and compute the mean and the variance of the pooled sample where the means are the same? Without going into details, the results would be as follows:

Mean radius (pooled): 10.01162592 and **Variance of radius (pooled): 0.083121563**

There is nothing surprising about this: after all, from our godlike perspective as creators of the example, we know perfectly well that the samples really are from the same population. Pooling the samples would be the same as simply taking a random sample of size 60 instead of two samples of size 30. The mean of the pooled data isn't much different from the two-sample means and the variance is pretty much like the separate variances.

But what if we alter the thought-experiment and imagine that the second sample actually comes from a different production line: one making 17" radius tires (7" bigger than the tiny 10" radius tires) for gas-guzzling monsters with macho supercharged engines and, ah, virile mag wheels? What would happen to the statistics we just computed?

Adding 7" to the existing Sample 2 to create a synthetic Sample 3 for our thought experiment produces the following results:

Figure 8-5. Sample statistics for tires including new Sample 3 modified by adding 7 to every value in Sample 2.

Samples	Mean Radius	Variance of radius
Sample 1	9.992921717	0.08067098
Sample 2	10.03033012	0.08771458
Sample 3	17.03033012	0.08771458

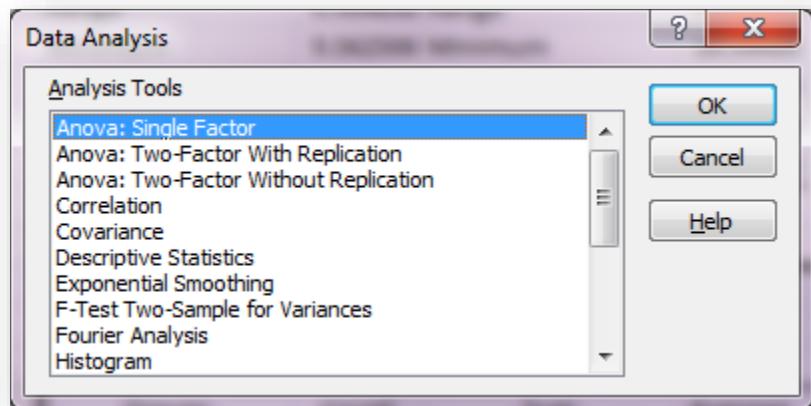
The mean of Sample 3 in Figure 8-5 is exactly 7" larger because we deliberately added 7" to every one of the observations we already had for Sample 2; it's no surprise that the mean increased by exactly 7". What may surprise you is that the variance of Sample 3 is exactly the same as the variance of Sample 2, but if you think about it, variance is computed as a function of the average deviations of the values in a set from their mean, so there was no increase in variability of the data after the artificial shift in the mean. We just shifted all the Sample 2 data up by 7"; we didn't increase the spread of the data at all in the new Sample 3.

Without doing a computation, we can say intuitively that *the average variance within the groups is the same regardless of the difference in the means of the two groups*. We won't delve too deeply just now into how to compute the average variance within the groups – that's coming in a few paragraphs.

But for now we compute the *overall variance of the pooled data* including Sample 1 and Sample 3: the mean is 13.51162592 and the variance is 12.67389725. The variance of the pooled data when the three samples have different means is bigger (~13.5) than when the two samples have the same mean (~0.08).

It turns out that the one-way analysis of variance (ANOVA) is based on something very similar to this kind of reasoning. After activating the **Data | Data Analysis** tools menu, we can use the function **Anova: Single Factor** (Figure 8-6).

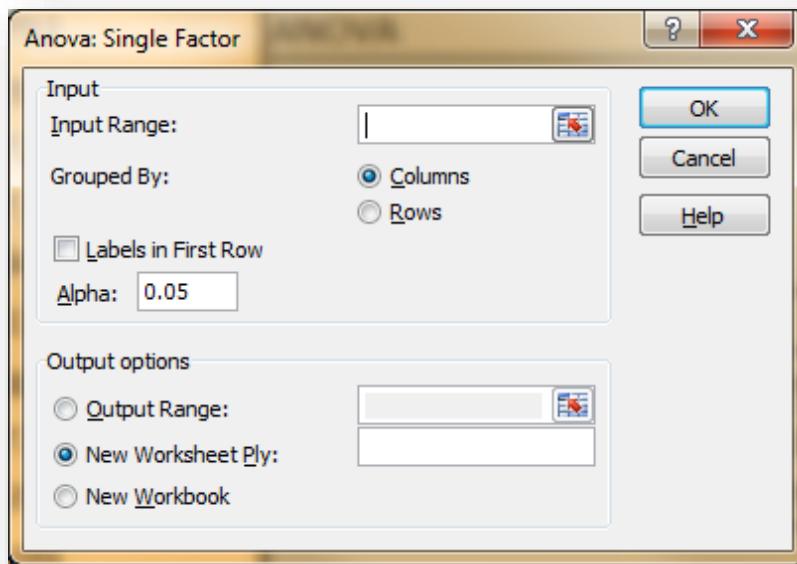
Figure 8-6. Choosing the ANOVA for testing differences of means classified by a single factor.



The concept of *single factor* refers to the classification of groups according to only one criterion; e.g., sample number or team or company or whatever we are interested in defining as our groups of interest.

Figure 8-7 shows the menu for entering the locations of the input data and the output in EXCEL 2010.

Figure 8-7. Menu for ANOVA single factor in Excel 2010.



Later, you'll study two-factor ANOVAs that classify groups by two criteria (factors) such as gender and geographical location, or company and size, or weight and height, or team and year....

Figure 8-8 shows the output for our example, starting with Samples 1 and 2 that were similar in their means:

Figure 8-8. ANOVA for tire data – samples 1 & 2 (means not different).

Anova: Single Factor				
SUMMARY				
Groups	Count	Sum	Average	Variance
Sample 1	30	299.7876515	9.992922	0.080671
Sample 2	30	300.9099037	10.03033	0.087715

ANOVA					
Source of Variation	SS	df	MS	F	P-value
Between Groups	0.020991	1	0.020991	0.249319	0.619444
Within Groups	4.883181	58	0.084193		
Total	4.904172	59			

Looking at Figure 8-8, you can see that the F-test in this case is the ratio of two variance-like measures of variability (called the *Mean Squares* and abbreviated *MS*).

- The Between Groups MS
 - Is a function of the differences in the means of the two (or more in general) samples,
 - Is in the row labeled *Between Groups* in the column marked *Source of Variation*
 - Is often printed as MS_{groups} in discussions of the results.
- The Within Groups MS
 - Is based on the average variations inside both samples
 - Is often written MS_{within}
 - Is often called MS_{error} .

The null and alternate hypotheses here are expressible in three ways that imply the same thing:

- $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$
- $H_0: MS_{groups} = MS_{within}$ $H_1: MS_{groups} \neq MS_{within}$
- $H_0: F \leq 1$ $H_1: F > 1$

As you can see in , the F-ratio of 0.249319⁹¹ is not statistically significant; there is no reason to reject the null hypothesis that the means of samples 1 and 2 are the same. The p-value of ~0.6 means that the chances of getting an F-ratio that large or larger with 1 and 58 degrees of freedom if the null hypothesis were true is about 3 out of 5 sampling experiments.

To check the calculation of F (just to help make sense of it), one can calculate =F.DIST.RT(0.249319,1,58) which sure enough gives 0.61944 just as the ANOVA table in Figure 8-8 shows.

⁹¹ Usually three significant figures are enough when publishing an F-test ratio, so we would normally adjust the number of decimal places to show 0.249.

Now let's look at the ANOVA for Samples 1 and 3:

Figure 8-9. ANOVA for tire data – samples 1 & 3 (different means created in demo data set).

Anova: Single Factor					
SUMMARY					
Groups	Count	Sum	Average	Variance	
Sample 1	30	299.7876515	9.992922	0.080671	
Sample 3	30	510.9099037	17.03033	0.087715	
ANOVA					
Source of Variation	SS	df	MS	F	P-value
Between Groups	742.8768	1	742.8768	8823.521	4.5E-65
Within Groups	4.883181	58	0.084193		
Total	747.7599	59			

Because we deliberately changed the mean of Sample 3 in this illustrative example,⁹² the MS_{groups} is huge: >742! However, because of the artificial nature of the modification, the MS_{within} is *exactly the same* as in Figure 8-8 comparing samples 1 and 2. The F-ratio ($MS_{groups} / MS_{within}$) is now enormous (>8823).

There is no reasonable doubt that we can *reject* the null hypothesis of equality of the means of samples 1 and 3 in this contrived example: the chances of getting such a large F-ratio (>8823) with 1 and 58 degrees of freedom are negligibly small (4.5E-65). The result is thus described as *extremely statistically significant*.

So let's recapitulate what you've seen in this section.

- ANOVA is based on the concept that if everything is the same in samples, a form of variance for individual samples versus the variance of the entire set of data should be similar.
- ANOVA depends strongly on the assumption that all samples being compared *have the same fundamental (parametric) variance*. We call this the assumption of *homoscedasticity*.⁹³

⁹² We added 7 to every value.

⁹³ Sokal & Rohlf write, “Synonyms for this condition are *homogeneity of variances* or **homoscedasticity**, a jawbreaker that makes students in any biometry class sit up and take notice. The term is coined from Greek roots meaning ‘equal scatter.’” (Sokal and Rohlf, Biometry: The Principles and Practice of Statistics in Biological Research 2012) p. 414.

8.8 The Model for Single-Factor ANOVA

The underlying model for this ANOVA can be written as follows:

$$Y_{ij} = \bar{\bar{Y}} + \alpha_i + \varepsilon_{ij}$$

where Y_{ij} is the j th observation in group i of the comparison;
 $\bar{\bar{Y}}$ is the overall (global) mean of all the observations;
 α_i is the average main effect of being in group i (i.e., $\bar{Y}_i - \bar{\bar{Y}}$);
 ε_{ij} is the residual variation for the j th observation in group i (i.e., $Y_{ij} - \bar{Y}_i$).

SS in the ANOVA table stands for *Sum of Squares* and, as you have seen already, df stands for *degrees of freedom*. A *Mean Square (MS)* is SS/df . The MS_G is the *Mean Square Among (or Between) the Groups*; for a groups, MS_G has $df_G = a - 1$ degrees of freedom. $MS_G = SS_G / df_G$. Similarly, MS_W is the *Mean Square Within Groups* (also known as the *Mean Square Error*, the *Residual Mean Square Error*, and the *Unexplained Mean Square Error* abbreviated MS_E) and has $df_W = \sum n - a$ where $\sum n$ is the sum of all the sample sizes of all the groups. The MS_W is a measure of the intrinsic variability of the system being studied; it is the variability left over after we have explained the variations in means among the groups caused by the main effects α_i .

The F-test for the presence of the main effects is

$$F_{[a-1], [\sum n - a]} = \frac{MS_G}{MS_W}$$

ANOVA is an immensely powerful tool that is extensible to many different situations. One of the simplest extensions of what you have already learned is to test for the presence of differences of the means of *more than two* samples at a time; however, the models can become much more sophisticated than the simple one-factor ANOVA introduced in this section. For example,

- We can compare measurements in the *same samples* with various *treatments*; for example, it is possible to use the paired-comparisons ANOVA to see if a measurement for each specific subject *before* administration of a medication is the same as the measurement for that subject after treatment. Such a comparison naturally reduces the MS_{within} and increases what we call the *power* of the test.
- It is possible to look at the effects of two different factors at the same time (two-way ANOVA) and even of many factors at once (multifactorial ANOVA). For example, we could study the effects of two different doping agents on the performance of integrated circuitry. In addition to telling if the different doping agents individually change performance, the two-way ANOVA would also reveal if the effect of one factor was different as a function of the other factor. We call such dependencies *interactions*.
- ANOVA with linear regression uses a model where an element of prediction based on the optimal least-squares regression coefficient to partition the variability into a component related to the regression and the residual variability assigned to the *Unexplained Mean Square Error*.

In all of these ANOVAs, the MS_W or *Unexplained Variability* or *Mean Square Error* (MS_E) tells us how much more work we have to do to understand the phenomenon we are studying; if the MS_E is large after we have explained part of the variability using models based on different classifications or regressions, then we still have mysteries to probe and to discover in the data.⁹⁴

⁹⁴ These terms (MS_W , MS_E and so on) are used interchangeably in applied statistics; different texts and journals have their preferred usage, but we should be ready to recognize all of them, so the text here is deliberately written with several versions of the terms.

8.9 Testing for the Equality of Two Means in an Investigation of Possible Dishonesty

There are other tests for evaluating the hypothesis that the means of two samples could be as different as they are (or more different) by chance (random sampling) alone. The exact choice depends on whether we already know the parametric variances of the populations from which the samples are drawn (in which case we use the Normal distribution to describe the distribution of differences between two sample means) or if we don't know the parametric variances and have to estimate them based on the sample variances (in which case we rely on Student's-t distribution).

By far the more common class of tests involves samples with unknown parametric variances when the sample variances are comparable.

Here's an example of a situation requiring a test for equality of means.

A computer systems administrator is investigating a possible case of insider fraud and needs to know if one of the employees has been spending a significantly longer time logged into the accounts payable database over the last three weeks compared with the previous three weeks. The reason is that the Chief Financial Officer has reported a series of unexplained discrepancies in the accounts payable files starting three weeks ago: there seem to be payments to nonexistent suppliers starting at that point. Suspicion has fallen on the Assistant Comptroller; the Chief Information Security Officer has asked for investigation of that employee's behavior. Has the suspect in fact significantly increased the length of his sessions?

The sysadmin finds the information from the log files shown in Figure 8-10.

The hypotheses are

$$\begin{array}{ll} H_0: \mu_1 \geq \mu_2 & \text{which is sometimes expressed } H_0: \mu_1 - \mu_2 \geq 0 \\ H_1: \mu_1 < \mu_2 & \text{which would be equivalent to } H_1: \mu_1 - \mu_2 < 0 \end{array}$$

Sometimes the symbol Δ (Greek capital delta) represents the hypothesized difference between the parametric means of the two populations from which the samples are drawn. When we are testing for the equality of the parametric means, $\Delta = \mu_1 - \mu_2 = 0$.

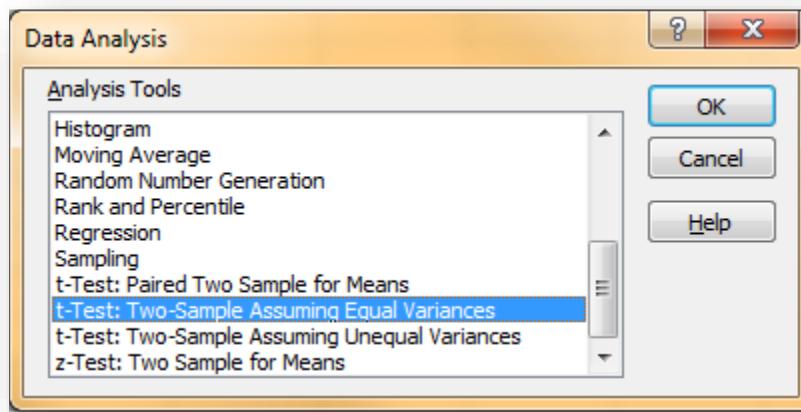
Figure 8-10. Log-file data on session length for suspected fraudster.

	A	B
1	Sample 1	Sample 2
2	18.9	30.49
3	16.89	26.63
4	15	29.69
5	22.74	25.03
6	11.58	28.56
7	20.91	29.67
8	19.04	31.29
9	17.23	24.97
10	25.69	24.86
11	18.94	25.5
12	19.29	34.33
13	21.78	33.87
14	17.58	28.55
15		32.33
16		22.7
17		23.49
18		31.05
19		29.91
20		31.26
21		29.84
22		29.7
23		29.25
24		26.91
25		30.2
26		22.68
27		22.5
28		33.74
29		30.49
30		36.11
31		24.67
32		25.54
33		30.81
34		26.9
35		27.53
36		28.7
37		31.03
38		21.99
39		25.51
40		36.7
41		32.19
42		35.49
43		30.77

8.10 T-Tests in Data Analysis

The easiest interface for a t-test is to use the Data | Data Analysis menu shown in Figure 8-11, which brings up a menu of options for t-tests.

Figure 8-11. Data Analysis tools for t-tests of means.



The last four functions shown in Figure 8-11 have the following purposes:

- **t-test: Paired Two Sample for Means** – used when the same subject is measured twice; e.g., to compare effect of a drug treatment by measuring blood hemoglobin concentration before and after administration of the dose on each subject rather than simply considering each group to be a random sample (discussed later in this course).
- **t-test: Two Sample Assuming Equal Variances** – two random samples with parametric variances thought to be identical.
- **t-test: Two Sample Assuming Unequal Variances** – two random samples with parametric variances thought to be different.
- **z-test: Two Sample for Means** – two random samples, each with a known parametric variance.

The first question is whether the observed sample variances are consistent with an assumption of homoscedasticity for the parametric variances.

The Data Analysis | Descriptive Statistics tool provides results which are shown slightly reformatted in Figure 8-12.

Next is to use the F-test for equality of two variances to test for homoscedasticity using the sample variances (12.560 and 14.732). As usual, we can pick the larger variance for the numerator and the smaller for the denominator so we can use the right-sided probability function given by =F.DIST.RT.

Figure 8-12. Descriptive statistics for two samples.

Statistic	Sample 1	Sample 2
Mean	18.890	28.891
Standard Error	0.983	0.592
Median	18.94	29.68
Mode	#N/A	30.490
Standard Deviation	3.544	3.838
Sample Variance	12.560	14.732
Kurtosis	0.928	-0.628
Skewness	-0.134	0.028
Range	14.110	14.710
Minimum	11.58	21.99
Maximum	25.69	36.70
Sum	245.57	1213.43
Count	13	42

The hypotheses for this test are

$$H_0: \sigma^2_1 = \sigma^2_2 \quad \text{and} \quad H_1: \sigma^2_1 \neq \sigma^2_2$$

Figure 8-13 shows the results. There is no reason to suspect that the parametric variances are unequal ($p = 0.401ns$), so we can choose the equal-variances version of the t-test for quality of means:

t-Test: Two Sample Assuming Equal Variances.

Figure 8-13. Test for equality of parametric variances in two samples.

CHECK FOR HOMOSCEDASTICITY:	Value	Formula
F-test for equal variances	1.173	=S7/R7
P(H0) for F-test	0.401	=F.DIST.RT(L18,S14-1,R14-1)
F-test result:	ns -- variances are equal	

Figure 8-14 shows the menu and entered data for the EXCEL t-Test: Two Sample Assuming Equal Variances data-analysis routine.

The results are unequivocal:

Because our null and alternative hypotheses are asymmetric

$$(H_0: \mu_1 \geq \mu_2 \text{ and } H_1: \mu_1 < \mu_2),$$

we use the *one-tail probability* (highlighted in bold in Figure 8-15) which is unquestionably extremely significant (1.555E11). There is virtually no question that the suspect has increased the length of his sessions significantly.

Figure 8-14. Data locations entered into menu for t-test.

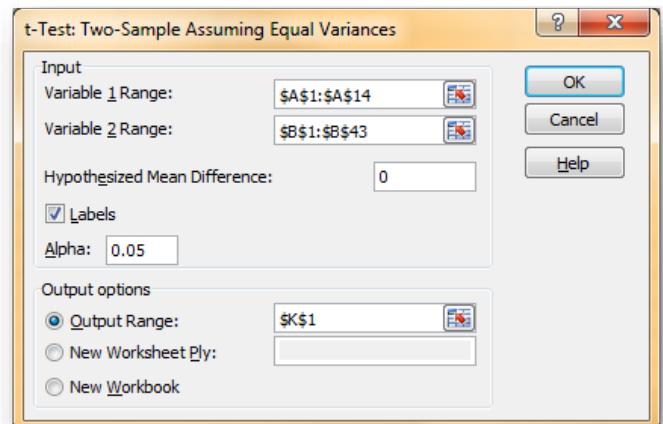


Figure 8-15. Results of t-test for equality of means assuming homoscedasticity.

t-Test: Two-Sample Assuming Equal Variances		
	Sample 1	Sample 2
Mean	18.89	28.89
Variance	12.56	14.73
Observations	13	42
Pooled Variance	14.24	
Hypothesized Mean Difference	0	
df	53	
t Stat	-8.350	
P(T<=t) one-tail	1.55E-11	
t Critical one-tail	1.674	
P(T<t) two-tail	3.10E-11	
t Critical two-tail	2.006	

Note that if we had been testing symmetrical hypotheses ($H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$), the appropriate probability would have been the $P(T \leq t)$ two-tail figure which, not surprisingly, is exactly twice (3.10E-11) the figure we chose (1.55E-11). That probability includes the likelihood that we could observe a difference of this magnitude or greater in either direction.

8.11 Critical Values

As a note of historical interest, the data analysis routine also provides *critical values* that demarcate the boundaries of the 0.05 “Alpha” stipulated in Figure 8-14.

The **t Critical one-tail** value is $t_{0.05|53} = 1.674$ and denotes the value of Student's t with

$$df = (n_1 + n_2 - 2) = (13 + 42 - 2) = 53$$

where 0.05 of the distribution lies to the left of this critical value. We can verify this assertion using

$$=T.DIST(-1.674, 53, 1) = 0.0500$$

as expected for the left tail. The same check could use

$$=T.DIST.RT(1.674, 53) = 0.0500$$

for the right tail..

Similarly, the **t Critical two-tail** value represents

$$t_{0.25|53} = 2.006$$

and demarcates 0.025 to the right of 2.006 and 0.025 to the left of -2.006

Again, verifying these assertions using EXCEL functions for the left and right tails, respectively,

$$=T.DIST(-2.006, 53, 1) = 0.0250$$

$$=T.DIST.RT(2.006, 53) = 0.0250$$

These *critical values* are holdovers from a time when statisticians did not have ready access to computers and relied on printed tables of probabilities for statistical decisions. Figure 8-16 shows such a table; users would look up the probability for a two-tailed critical value along the top and go down the column for the critical value for degrees of freedom.⁹⁵

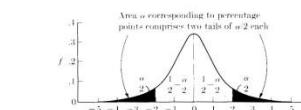
Today, few statisticians use statistical tables unless they also like to use tables of logarithms, or to perform statistical calculations by hand or using a calculator, or to use overhead projectors and photocopies on acetates for prepared images in lectures instead of computer programs and LCD projectors, or to chisel cuneiform symbols on clay tablets instead of using email.

However, if you are stranded on Pluto without access to your cerebral computational brain implant, having a set of statistical tables may allow you to check a published calculation or two.⁹⁶

Figure 8-16. Image of a statistical table from a 1981 set of tables.

TABLE 12 Critical values of Student's t -distribution.

α	0.9	0.5	0.4	0.2	0.1	0.05	0.02	0.01	0.001	α	P
1	.158	1.000	1.376	3.078	6.314	12.706	31.821	63.657	636.619	1	
2	.142	.816	1.061	1.886	2.920	4.303	9.965	9.925	31.598	2	
3	.137	.765	.978	1.638	2.353	3.182	4.541	5.841	12.924	3	
4	.134	.727	.913	1.476	2.105	2.763	3.729	5.787	10.010	4	
5	.132	.707	.896	1.476	2.015	2.571	3.365	4.032	6.869	5	
6	.131	.718	.906	1.440	1.943	2.447	3.143	3.707	5.959	6	
7	.130	.711	.896	1.415	1.894	2.365	2.998	3.499	5.408	7	
8	.130	.706	.889	1.397	1.864	2.306	2.896	3.355	5.041	8	
9	.130	.703	.883	1.381	1.839	2.282	2.857	3.311	4.741	9	
10	.129	.700	.879	1.372	1.812	2.228	2.764	3.169	4.587	10	
11	.129	.697	.876	1.363	1.794	2.201	2.718	3.106	4.437	11	
12	.128	.695	.873	1.356	1.782	2.179	2.681	3.055	4.318	12	
13	.128	.693	.871	1.350	1.770	2.160	2.650	3.026	4.211	13	
14	.128	.692	.868	1.345	1.761	2.145	2.624	2.977	4.140	14	
15	.128	.691	.866	1.341	1.753	2.131	2.601	2.947	4.073	15	
16	.128	.690	.865	1.337	1.746	2.120	2.583	2.921	4.015	16	
17	.128	.689	.863	1.333	1.740	2.110	2.567	2.894	3.965	17	
18	.127	.688	.861	1.329	1.734	2.091	2.547	2.861	3.922	18	
19	.127	.688	.861	1.328	1.729	2.093	2.539	2.861	3.883	19	
20	.127	.687	.860	1.325	1.725	2.086	2.530	2.845	3.850	20	
21	.127	.686	.859	1.323	1.721	2.080	2.518	2.831	3.819	21	
22	.127	.686	.858	1.321	1.717	2.074	2.502	2.819	3.792	22	
23	.127	.685	.858	1.319	1.714	2.069	2.486	2.803	3.763	23	
24	.127	.685	.857	1.318	1.711	2.064	2.462	2.797	3.745	24	
25	.127	.684	.856	1.316	1.708	2.060	2.448	2.787	3.727	25	
26	.127	.684	.856	1.315	1.706	2.056	2.437	2.779	3.707	26	
27	.127	.684	.855	1.314	1.704	2.052	2.427	2.769	3.687	27	
28	.127	.683	.855	1.313	1.701	2.048	2.417	2.763	3.674	28	
29	.127	.683	.854	1.313	1.699	2.045	2.402	2.756	3.659	29	
30	.127	.683	.854	1.310	1.697	2.042	2.397	2.750	3.646	30	
40	.126	.681	.851	1.303	1.684	2.031	2.423	2.704	3.551	40	
60	.126	.679	.848	1.297	1.671	2.004	2.403	2.680	3.460	60	
120	.126	.677	.845	1.289	1.658	1.980	2.358	2.617	3.373	120	
∞	.126	.674	.842	1.282	1.645	1.960	2.326	2.576	3.291	∞	



⁹⁵ (Rohlf and Sokal 1981) p 81. Used with kind permission of author F. J. Rohlf.

⁹⁶ Geeks should note that carrying statistical tables around will *not* increase your attractiveness to potential dates.

8.12 ANOVA: Single Factor vs T-test for Equality of Means

The Student's t-test for equality of means produces results consistent with those of an ANOVA. To demonstrate this similarity, Figure 8-17 shows the ANOVA using EXCEL's Data Analysis routine Anova: Single Factor for the same data as the t-test just discussed.

Students should note the equality of the

Figure 8-18. ANOVA for same data as t-test example.

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Sample 1	13	245.57	18.89	12.56		
Sample 2	42	1213.43	28.89	14.73		

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	992.9637	1	992.9637	69.7	3.10E-11	4.02
Within Groups	754.7354	53	14.2403			
Total	1747.699091	54				

$P(T \leq t)$ two-tail = 3.10E-11 from the t-test and the

P-value = 3.10E-11 in the ANOVA for the $F[1,53] = 69.7^{***}$.

ANOVA of this kind always performs a *two-tailed* F-test for the H_0 of *equality* of means.

In the t-test and in the ANOVA, the null hypothesis is rejected: the difference between the means is extremely statistically significant at the 0.001 level (there is essentially no realistic chance of encountering a difference this large or larger by chance alone if the null hypothesis of equality is true) and therefore we accept the alternate hypotheses:

- In the original t-test shown in § 8.10, that the mean session length is *longer* in the second sample than in the first sample (a *one-tailed* test);
- In the ANOVA, that the mean session length is *different* between the samples (a *two-tailed* test). Given that the observed difference is that $\bar{Y}_2 > \bar{Y}_1$, we come to the same conclusion as in the t-test: the Assistant Comptroller has indeed been spending a lot more time logged into the accounts payable database in these last three weeks than in the three weeks before. Maybe it's time to do some more forensic research on the case.

8.13 Testing for Equality of Means Given Parametric Mean and Parametric Standard Deviation v Sample Mean

Sometimes we have such a large amount of historical information about the mean and standard deviation of a measurement that we are willing to consider those *parametric* statistics. In §5.7, we discussed how to compute *confidence limits* based on parametric mean and parametric standard deviation. The same calculations can be adapted to test the hypothesis that an observed *sample* mean and observed *sample* standard deviation are consistent with the null hypothesis. Here are screenshots of the formulas and the calculated values from a homework exercise in the QM213 course at Norwich University.

Figure 8-19. Using parametric mean & standard deviation for an hypothesis test about a sample mean.

A	B	C	D	E	F	G	H	I
2								
3								
4	a.	$H_0: \mu \leq 15$	15	$= \mu$				
5		$H_a: \mu > 15$		This is a one-tailed test.				
6								
7		$n = 35$	for sample of calls drawn at random					
8		$\bar{X} = 17$	average length of surveys in sample					
9		$\sigma = 4$	parametric standard deviation from mass of prior data					
10		$\alpha = 0.01$	significance level					
11								
12		$se(\bar{X}) = \sigma/\sqrt{n} =$	$= +C9/SQRT(C7)$					
13								
14	b.	$z = (\bar{X} - \mu)/(\sigma/\sqrt{n}) =$	$= +(C8-D4)/D12$					
15								
16	c.	Upper-tail p-value is area to the right of the test statistic						
17		Therefore we use 1-NORM.S.DIST function in Excel						
18		$p = 1 - NORM.S.DIST(D14, 1)$	**					
19								
20	d.	$p \leq \alpha$	therefore	reject H_0 & charge premium				
21								

9.4a Fowle Marketing Telephone Surveys

Charges by a call center are based on mean survey time of 15 minutes or less. They charge a premium rate if the survey takes longer than 15 minutes on average. Do the data below support premium charges when using a significance level of 0.01? Do the data below support premium charges when using a significance level of 0.01?

a.	$H_0: \mu \leq 15$	15	$= \mu$					
	$H_a: \mu > 15$		This is a one-tailed test.					
	$n = 35$	for sample of calls drawn at random						
	$\bar{X} = 17$	average length of surveys in sample						
	$\sigma = 4$	parametric standard deviation from mass of prior data						
	$\alpha = 0.01$	significance level						
	$se(\bar{X}) = \sigma/\sqrt{n} =$	0.6761						
b.	$z = (\bar{X} - \mu)/(\sigma/\sqrt{n}) =$	2.9580						
c.	Upper-tail p-value is area to the right of the test statistic							
	Therefore we use 1-NORM.S.DIST function in Excel							
	$p = 0.001548$	**						
d.	$p \leq \alpha$	therefore	reject H_0 & charge premium					

8.14 Computing a t-test for Equality of Means without Raw Data

Finally, in case you need it, here is the intimidating general formula for the t-test for equality of two means if you *don't* have access to the raw data but only the means, the sample variances, and the sample sizes. The symbol Δ is just the hypothesized difference between the parametric means – usually 0.

$$t_{df} = \frac{(\bar{Y}_1 - \bar{Y}_2 - \Delta)}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}} \quad \text{where} \quad df = \frac{\left(\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2} \right)^2}{\frac{1}{n_1-1} \left(\frac{s^2_1}{n_1} \right)^2 + \frac{1}{n_2-1} \left(\frac{s^2_2}{n_2} \right)^2}$$

Aren't you glad you use EXCEL? Although there is no instant function for these calculations, as long as you label all the steps of your calculations, you will have no problem computing the values step by step and then using the =T.DIST.2T function to compute the probability of observed t-value if the null hypothesis were true.

8.15 The T.TEST Function

For completeness, let's look at another t-test for the equality of means, this time using raw data and the EXCEL =T.TEST function, which requires the array (both columns of data but NOT the headings), the number of tails (1 or 2), and the choice of model:

- **Type = 1 for paired comparisons** in which the same entities are observed under two different circumstances; e.g., if the same glue job were repeated with standard application glue and then with molecular transubstantiation glue. There must, naturally, be exactly the same number of observations if we are doing a paired comparison. Another typical example is a before/after comparison of individual test subjects; e.g., a paired comparison of *each person's* weight before and after a month of a special diet.
- **Type = 2 for two samples** (as in our example of the possibly corrupt Assistant Comptroller) where we assume that the parametric variances are equal (homoscedasticity). In our case, lacking any reason to do otherwise, we choose this assumption.
- **Type = 3 for two heteroscedastic samples** (where we assume that the parametric variances are unequal).

Using the same data as in the t-test of § 8.9 *et seq.*,

=T.TEST(A2:A14,B2:B43,1,2) = 1.55E-11

Which is exactly the same as the result of the one-tailed t-test in §8.10 (Figure 8-15).

Similarly,

=T.TEST(A2:A14,B2:B43,2,2) = 3.10E-11

exactly as in both the two-tailed t-test and in the ANOVA of §8.12 (and Figure 8-17).

9 Analyzing Relationships Among Variables

9.1 Introduction to Analyzing Relations

There are many cases where we need to discuss more than one aspect of the entities we are interested in understanding. For example, in analyzing productivity figures in a factory as part of an operations research study, we may collect information about several aspects of each product line beyond just the daily production such as

- **Day of the week** – maybe production changes systematically during the work week; keeping track of exact dates and therefore allowing us to track Mondays, Tuesdays and so on could be useful.
- **Shift supervisor** – perhaps particular supervisors differ in their effects on productivity; a nasty supervisor, for example, might cause resentment among workers and get poorer productivity than a good supervisor. Alternatively, perhaps lower production in “Bob’s” shift is due to theft orchestrated by Bob!
- **Type of equipment on each production line** – maybe specific lines are affected by differences in the machinery.
- **Ambient temperature in factory** – perhaps differences in productivity can be traced to working conditions such as excessive heat during afternoon shifts compared with nighttime or morning shifts.

Figure 9-1 shows some sample data with a few of these observations. There is no theoretical limit on the level of detail that we can collect in scientific and professional studies. What limits our ability to study all aspects of reality are such factors as

- **The difficulty or impossibility of quantifying specific attributes of reality;** e.g., there is no easy, simple measure of such human attributes as *honesty* or *originality*; and there is no immediate, simple measure of a product's *utility* or *marketing appeal*.
- **Ability to define metrics (ways of measuring something);** e.g., marketing appeal might be measured through studies of purchasing habits for that product.
- **The complexity of acquiring the data;** e.g., a measure that we think might be indicative of *honesty* might require extensive testing of every subject, possibly in different situations and environments. Similarly, measuring a product's *utility* might involve extensive studies of how consumers actually use the product over a period of years.
- **Ability to identify independent factors possibly affecting the variable(s) of interest;** e.g., in a wide range of different populations grouped by such factors as age, gender, socio-economic status, and so on.
- **The controllability of factors;** e.g., it might be possible to impose experimental conditions on subjects in a study of honesty by giving them tasks in a laboratory or via computer; it might be much more difficult to perform such studies in the real world.
- **Increased costs resulting from increased complexity of data gathering:** It's cheaper to weigh bolts to see if they vary in weight than it is to study the marketing benefits of offering five different shades of those bolts.

Such considerations lead to *multifactorial analysis*. Simple approaches to handling such multifactorial data include *cross-tabulations* (also known as *contingency tables*), *scatterplots* for showing two variables at a time, *correlation coefficients* to express the intensity of a relationship between two variables, and *regression equations* to predict a variable's value as a function of one or more other variables.

Figure 9-1. Multiple variables in observations about production line (first week of data only).

Date	DoW	Line	Supervisor	TOTAL
2018-08-06	MON	A	Alice	33855
	MON	B	Alice	33114
	MON	C	Bob	19708
	MON	D	Bob	17834
2018-08-07	TUE	A	Bob	11116
	TUE	B	Alice	42171
	TUE	C	Bob	18597
	TUE	D	Alice	47431
2018-08-08	WED	A	Alice	35004
	WED	B	Bob	14658
	WED	C	Alice	44574
	WED	D	Bob	24398
2018-08-09	THU	A	Charlie	24297
	THU	B	Alice	44146
	THU	C	Darlene	52724
	THU	D	Bob	25500
2018-08-10	FRI	A	Darlene	32240
	FRI	B	Alice	39517
	FRI	C	Bob	21210
	FRI	D	Charlie	30798
2018-08-11	SAT	A	Charlie	20321
	SAT	B	Darlene	28795
	SAT	C	Charlie	25166
	SAT	D	Darlene	46736
2018-08-12	SUN	A	Charlie	12255
	SUN	B	Darlene	37164
	SUN	C	Charlie	25285
	SUN	D	Darlene	52654



9.2 Cross-Tabulations (Contingency Tables)

Let's consider the production-line example introduced above. Figure 9-2 shows the first week of collected data, with information on the day of week (Monday through Sunday), production line (A, B, C, or D) and supervisors (Alice, Bob, Charlie and Darlene). As you can see, it's difficult to make sense of these unaggregated data.

One summary representation, shown in Figure 9-2 is three-way cross-tabulation (or *three-way contingency table*) that shows production classified by three of the variables: it collapses the data by combining the different dates and adding the production values by *day of the week*, *production line* (A, B, C, D) and *supervisor*. The table also provides subtotals.

The table in Figure 9-2 was generated using the EXCEL Insert | PivotTable function accessible (Figure 9-3).

Figure 9-2. Cross-tabulation of production data showing breakdown by Day of Week and Supervisor.

Sum of TOTAL	Column Labels				
	Alice	Bob	Charlie	Darlene	
Row Labels					Grand Total
SUN			37,540	89,818	127,358
	A		12,255		12,255
	B			37,164	37,164
	C		25,285		25,285
MON	66,969	37,542			104,511
	A	33,855			33,855
	B	33,114			33,114
	C		19,708		19,708
TUE	89,602	29,713			119,315
	A		11,116		11,116
	B	42,171			42,171
	C		18,597		18,597
WED	79,578	39,056			118,634
	A	35,004			35,004
	B		14,658		14,658
	C	44,574			44,574
THU	44,146	25,500	24,297	52,724	146,667
	A		24,297		24,297
	B	44,146			44,146
	C			52,724	52,724
FRI	39,517	21,210	30,798	32,240	123,765
	A			32,240	32,240
	B	39,517	21,210		39,517
	C		30,798		21,210
SAT			45,487	75,531	121,018
	A		20,321		20,321
	B			28,795	28,795
	C		25,166		25,166
Grand Total	319,812	153,021	138,122	250,313	861,268

Figure 9-3. Insert | PivotTable symbol.

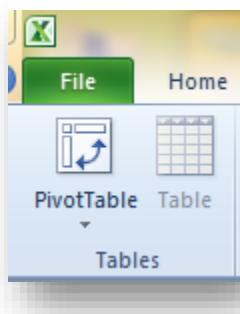
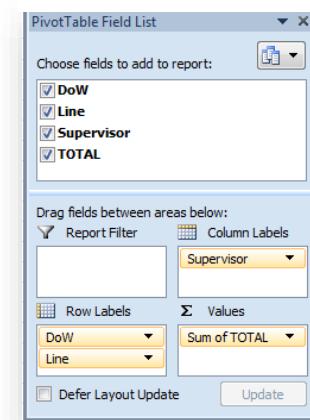


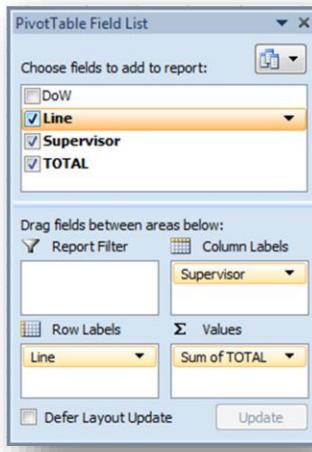
Figure 9-2 was generated using the options shown in the PivotTable Field List (Figure 9-4).

Figure 9-4. PivotTable settings for three-way cross-tabulation.



Another representation summarizes production by production line and by supervisor. Figure 9-5 shows the PivotTable settings for this arrangement and **Figure 9-6** shows the resulting table.

Figure 9-5. PivotTable settings for two-way contingency table.



This table (**Figure 9-6**) is a *two-way* contingency table because it shows counts classified by two variables: *production line* and *supervisor*. The table includes subtotals by those variables.

Figure 9-6. Two-way contingency table for production line and supervisor.

Row Labels	Sum of TOTAL	Column Labels				Grand Total
		Alice	Bob	Charlie	Darlene	
A	68,859	11,116	56,873	32,240	169,088	
B	158,948	14,658		65,959	239,565	
C	44,574	59,515	50,451	52,724	207,264	
D	47,431	67,732	30,798	99,390	245,351	
Grand Total	319,812	153,021	138,122	250,313	861,268	

9.3 Filtering Data for Temporary Views

Another useful EXCEL tool for working with complex tables is the **Filter** function accessible through the **Sort & Filter** drop-down menu (Figure 9-7). Clicking on the Filter option inserts pull-down menus in the columns of the table that one has highlighted.

Figure 9-7. Sort & Filter menu.

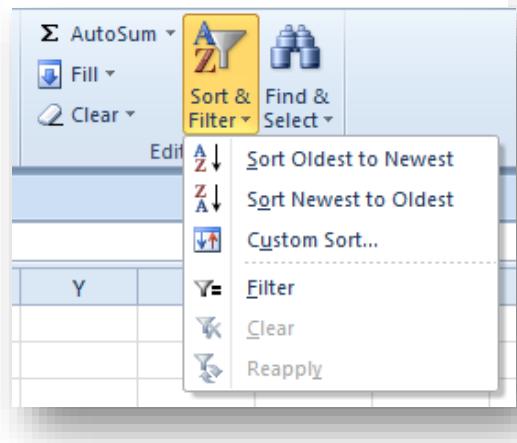


Figure 9-9 shows the pull-down menu:

Figure 9-9. Pull-down menu for DoW column.

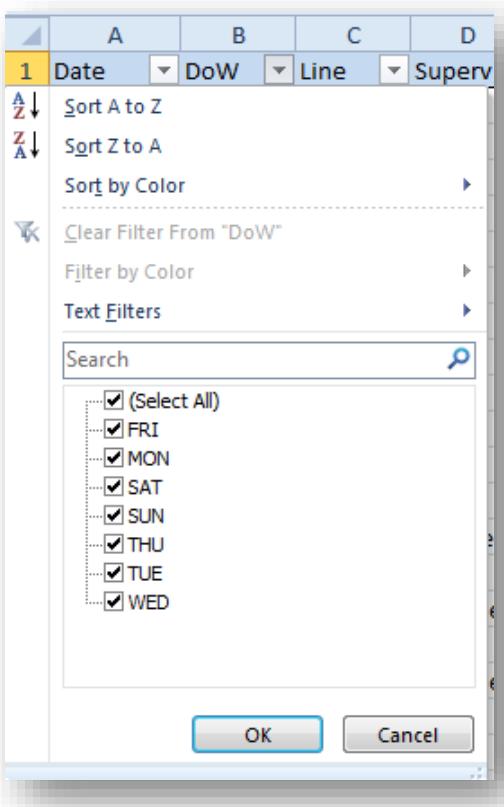


Figure 9-8 shows the headings and a few rows resulting from highlighting the entire table shown in Figure 9-1 at the start of §9.2 and selecting the **Filter** function.

Figure 9-8. Heading row showing pull-down menu tabs.

Date	DoW	Line	Superv	TOTAL
2018-08-06	MON	A	Alice	33,855
	MON	B	Alice	33,114
	MON	C	Bob	19,708
	MON	D	Bob	17,834
2018-08-07	TUE	A	Bob	11,116

Figure 9-10 shows the results of clicking on (Select All) to remove all the check marks and then clicking only on the SAT and SUN checkboxes.

Figure 9-10. Filtered subset of table showing only SAT and SUN data.

Date	DoW	Line	Superv	TOTAL
2018-08-11	SAT	A	Charlie	20,321
	SAT	B	Darlene	28,795
	SAT	C	Charlie	25,166
	SAT	D	Darlene	46,736
2018-08-12	SUN	A	Charlie	12,255
	SUN	B	Darlene	37,164
	SUN	C	Charlie	25,285
	SUN	D	Darlene	52,654

9.4 Charts for Contingency Tables

Two-way contingency tables can be represented graphically using a variety of charts. A popular graph type is the vertical bar chart with one variable represented by position along the abscissa and the other by the position (and usually color) of a bar in a series clustered over each value of the abscissa variable. In Figure 9-11, the abscissa shows the days of the week and the clustered bars represent the production lines. Such charts are easily created in EXCEL.

Figure 9-11. Clustered bar chart showing production totals for day of week and production line

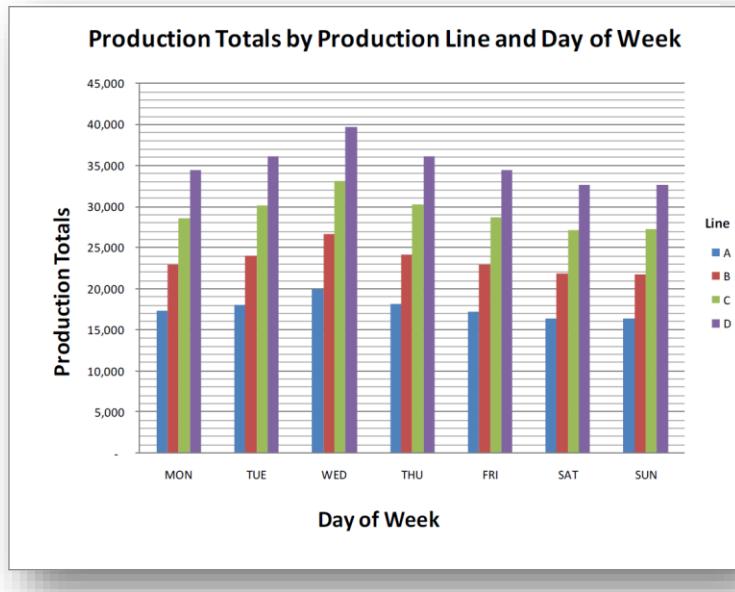
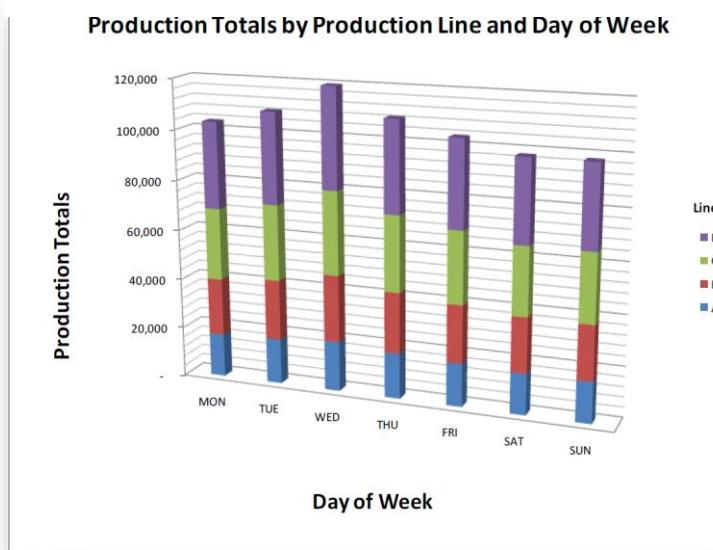


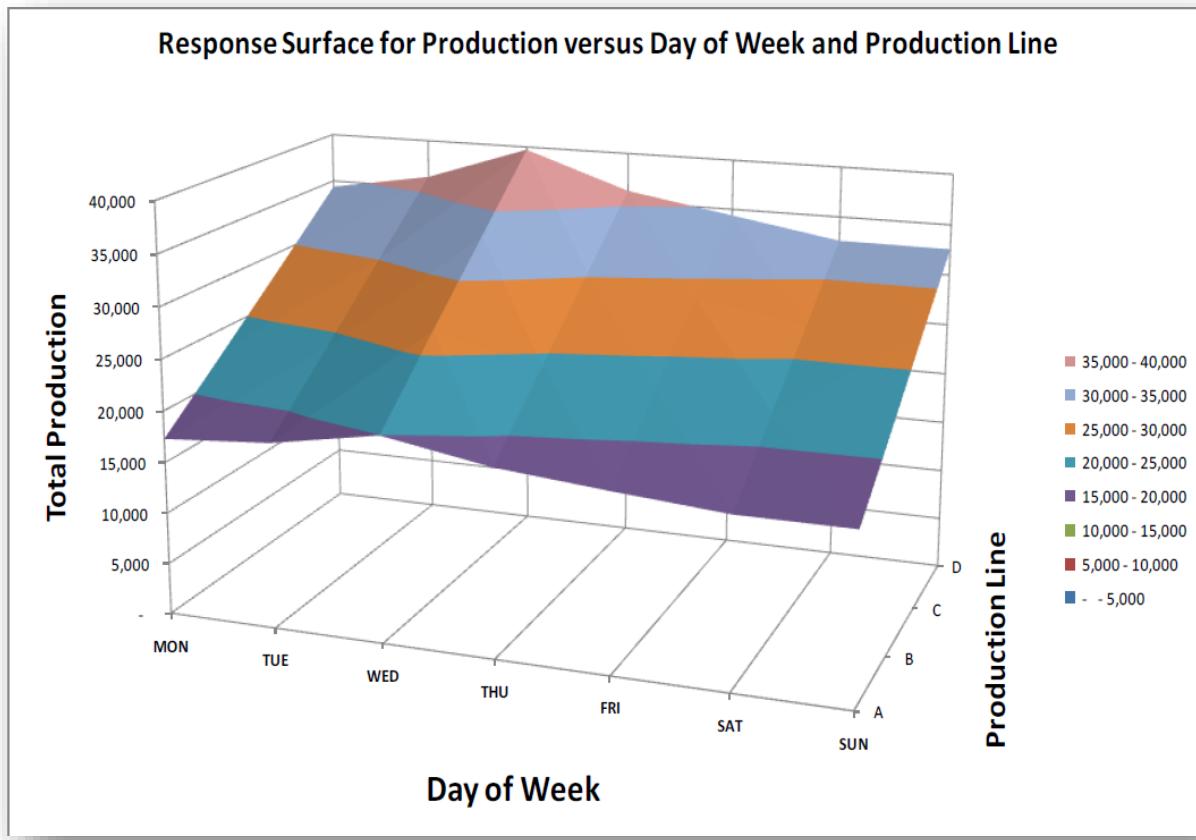
Figure 9-12. Stacked bar chart for production totals by day of week.

Another version of these data is the stacked bar chart (Figure 9-12), in which the second variable has portions of a single bar representing the total for all of its values. This particular example also uses features of EXCEL such as the ability to tilt and rotate three-dimensional graphs.



Another useful representation is the response surface, which shows a two-dimensional surface combining information from two variables versus a third. Figure 9-13 is a response surface for production versus both day of week and production line. Such charts can easily be prepared in EXCEL and there are many options for enhancing their appearance. In particular, one can rotate the image on any axis to clarify relationships that are of interest.

Figure 9-13. Three-variable response-surface chart.



9.5 Scatterplots and the Intuitive Grasp of Relationships

Often we measure two quantitative variables for the same entities and want to see how the data fall out in general terms. For example, Figure 9-14 shows data about the rating of a new advertising campaign by people of different ages as part of a marketing study to see if the advertisements appeal differently to viewers as a function of their age.

Figure 9-15. Scatterplot of rating vs age of respondent.

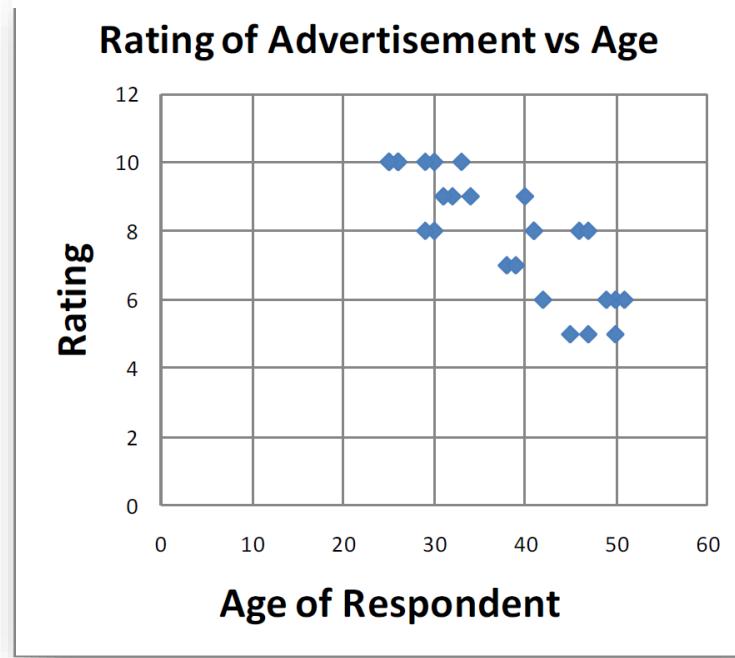


Figure 9-14. Data on responses to ads and age of respondent.

Consumer age	Rating of Advertisement
42	6
25	10
50	6
30	10
32	9
25	10
26	10
31	9
46	8
50	5
45	5
33	10
41	8
29	10
30	8
38	7
34	9
47	5
49	6
26	10
29	8
26	10
41	8
47	8
39	7
40	9
51	6
25	10

Figure 9-15 shows the scatterplot derived from these marketing-research data. Each point represents the particular combination of consumer age and rating of the advertisements for a specific respondent. These figures are easy to create in EXCEL.

At an intuitive level, a scatterplot in which the dots form a *slanted, tight* cloud naturally suggests to even a naïve viewer that perhaps there is some sort of relationship between the two variables shown. In Figure 9-15, for example, the immediate impression is that perhaps the ratings fall as the age of the respondent rises.

Contrariwise, a formless, wide cloud of dots in a scattergram suggests that perhaps there is no relation between the variables. However, without rigorous analysis, such impressions remain just that – impressions rather than actionable conclusions. For real-world decisions, we need to have quantitative estimates of the strength of the relationships and the probability that we are seeing the results of raw chance in the random sampling.

Proper statistical analyses of these impressions can involve *correlation* and *regression*, depending on the assumptions of the analysis. Correlation gives us an estimate of the *strength* of the relationship, if any; regression gives us estimates of the precise quantitative nature of the relationship, if any.

9.6 Pearson Product-Moment Correlation Coefficient, r

The intensity of the relationship between two variables that are both measured independently, and for which neither is viewed as a predictor of the other, is measured by an important statistic called the *Pearson product-moment correlation coefficient, r* .⁹⁷ This statistic applies to Normally distributed interval scales; there are other forms of correlation coefficient suitable for ordinal scales.

The notion of *independence* is important because it can determine whether to represent the relation between two variables in a data set in terms of the *correlation coefficient*, which implies no predictive rule and treats both variables as equally free to vary, in contrast with the *regression coefficient* (discussed in the next section) which usually assumes that one of the variables (the *independent variable*) can be used as a predictor of the other (the *dependent variable*).

In the example shown in Figure 9-14 and Figure 9-15, on the previous page, we collected data about the *age of respondents* and about their *rating of an advertisement*. Intuitively, we wouldn't expect anyone to be interested in predicting the age of a respondent by asking them how they feel about an ad; however, it could mean a lot to be able to measure the strength of relationship between how people feel about an ad based on their age if one wants to reach a particular demographic slice of the population. Judging from the scattergram in Figure 9-15, it looks roughly as if the older people liked the ad being tested less than the younger people. The natural, spontaneous tendency is to put the age on the abscissa and the response to the ad on the ordinate; that's a typical arrangement: the independent variable goes on the X-axis and the dependent variable goes on the Y-axis. Reversing the two would look funny in this case, although not necessarily in other cases.

Consider now a study of the responses to two different advertisements shown in Figure 9-16. Each vertical pair (e.g., 1, 1) represents the rating by an individual of their response to each of two ads.

Figure 9-16. Ratings of two different ads by 28 people.

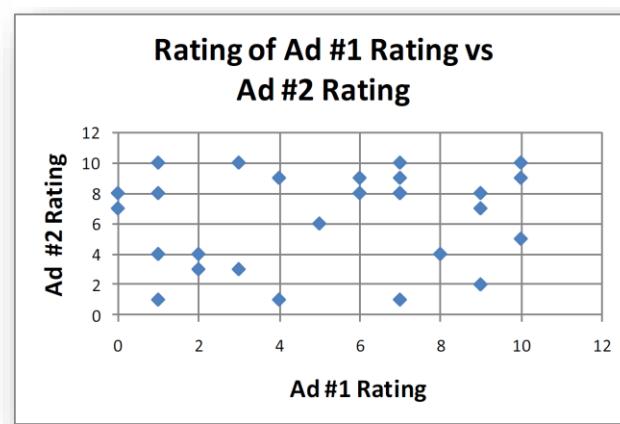
Rating of Ad #1	1	3	10	2	3	2	7	9	7	10	4	9	5	0	6	7	10	7	8	1	1	7	1	4	9	0	6	7
Rating of Ad #2	1	3	10	3	10	4	8	8	8	9	1	7	6	7	8	9	5	1	4	4	8	10	10	9	2	8	9	8

In this case, there's no particular reason why we would assign priority to one ad or the other; both are what we call *independent variables*. We can still be interested in the intensity of the relationship – the association – between these responses, but we don't normally think in terms of predicting one from the other.

Figure 9-17 shows a scatterplot for the responses to the two ads. The artificial data in this example were constructed to be randomly related – and the scatterplot shows a typical appearance for such data, with no obvious pattern.

The computation of the product-moment correlation coefficient, r , is best left to the statistical package.

Figure 9-17. Scatterplot for paired responses to two ads.



⁹⁷ Karl Pearson (1857-1936) was one of the founders of modern statistics. (O'Connor and Robertson 2003)

EXCEL has a function for computing r : =CORREL(array1, array2) which instantly provides the coefficient. In our current example, $r = 0.0.156246 \approx 0.156$.

Notice that in this case, it doesn't matter which variable is selected for which axis; we say that these are *two independent variables*.

The correlation coefficient r has properties that are easy to understand:

- Two variables with *no relationship* between them at all have a correlation coefficient $r = 0$.
- Two variables in which a larger value of one variable is *perfectly* associated with a correspondingly larger value of the other have an $r = +1$. E.g., if we calculate r for the height of individuals measured in inches and then measured in centimeters, the data should have $r = 1$ because knowing one measurement should allow computation of the other measurement without error.
- If a larger value of one variable is *perfectly* associated with a systematically smaller value of the other, the $r = -1$. For example, imagine giving children bags of candies containing exactly 20 candies to start with; in this silly situation, calculating the r for the number of candies eaten and the number of candies left should produce $r = -1$ because the more candies are eaten, the fewer are left – and there should be zero error in the data.
- In both of these extremes, knowing the value of one of the variables allows perfect computation of the value of the other variable. We say that there is no unexplained error in the relationship. However, if we introduce errors, r will decline from $+1$ or increase from -1 . For example, if we measure the height of a subject with a laser system correct to 0.01cm but measure the height in inches using a piece of string with only whole numbers of inches marked on the string, it's likely that the correlation coefficient will be less than perfect. Similarly, if a three-year-old is counting the candies that are left (and perhaps eating some surreptitiously), the data for the candy correlation may very well produce an $r > -1$.

9.7 Computing the Correlation Coefficient Using EXCEL

As mentioned in the previous section, the function =CORREL(array1, array2) mentioned in the previous section computes r with no additional information and no options.

However, the **Data Analysis** function

Correlation offers more options than the simple function when we need to compute r for more than one pair of variables. It's especially useful because it calculates correlations for all possible pairs of the data.

For example, imagine that a study (Figure 9-18) of the percentage of foreign outsourcing for ten companies is tested for a possible correlation with the frequency of industrial espionage expressed in relation to the number of employees in the company. In addition, the researchers also record the percentage profitability over the past decade for each company.⁹⁸

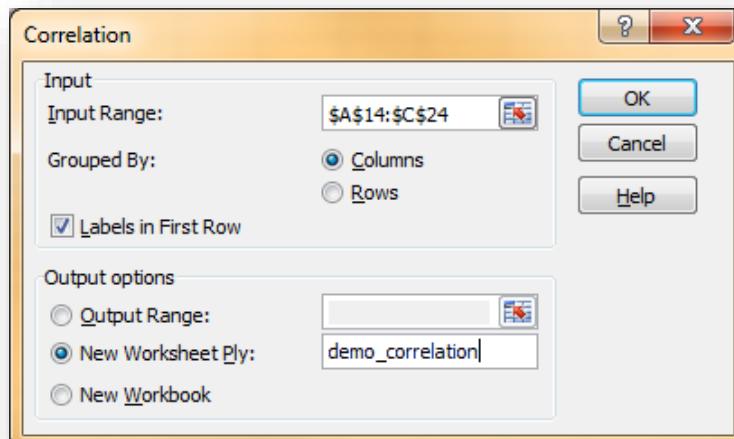
Figure 9-18. Outsourcing, espionage and profitability.

Percentage Overseas Outsourcing	Frequency of Industrial Espionage/1000 employees	Profitability over Last 10 Years
94%	9	-8%
18%	7	14%
54%	7	4%
36%	4	11%
54%	5	24%
95%	10	-2%
39%	9	2%
41%	8	9%
41%	7	7%
91%	8	2%

⁹⁸ These are *made-up figures*; do not interpret them as indicating anything real – they are created only for the purpose of illustration in this text.

Using the Correlation function from the Data Analysis menu (Figure 9-19),

Figure 9-19. Entering ranges and options into Correlation menu.



we generate a simple table of results showing the correlation coefficients r for each pair of variables in this made-up example. The default output lacks effective formatting, as you can see in Figure 9-20, where the columns are too narrow to display the column headings.

Figure 9-20. Unformatted output from Data Analysis | Correlation tool.

Percentage Overseas Espionage over Last 10 Years			
Percentag	1		
Frequency	0.5326	1	
Profitabili	-0.62706	-0.76122	1

However, some simple formatting (e.g., highlighting the columns and double-clicking any of the column boundaries – and the same for the heading row) produces a more usable report (Figure 9-21).

Figure 9-21. More readable formatted output from Data Analysis | Correlation tool.

CORRELATION COEFFICIENTS	Percentage Overseas Outsourcing	Frequency of Industrial Espionage/1000 employees	Profitability over Last 10 Years
Percentage Overseas Outsourcing	1		
Frequency of Industrial Espionage/1000 employees	0.533	1	
Profitability over Last 10 Years	-0.627	-0.761	1

For example, the correlation coefficient r for *Percentages Overseas Outsourcing* and *Frequency of Industrial Espionage/ 1000 Employees* is 0.533 and r for the *Outsourcing percentage* and *10-year profitability* is about -0.63.

9.8 Testing the Significance of the Correlation Coefficient

How do we know if a sample's correlation coefficient (r) is consistent with the hypothesis that the parametric correlation (ρ)⁹⁹ has a particular value? As usual, it's possible to compute a test statistic based on a correlation coefficient (r) based on a sample of size n that is distributed as a Student's-t statistic:

$$t_{[v]} = (r - \rho) / s_r$$

where

$t_{[v]}$ = the test statistic with $v = n - 2$ degrees of freedom

s_r = standard error of the correlation coefficient:

$$s_r = \sqrt{(1 - r^2)/(n - 2)}$$

Thus

$$t_{(n-2)} = (r - \rho) / \sqrt{(1 - r^2)/(n - 2)}$$

If our hypotheses are $H_0: \rho = 0$ and $H_1: \rho \neq 0$, then

$$t_{(n-2)} = r / \sqrt{(1 - r^2)/(n - 2)}$$

In the example discussed in §9.7, $n = 10$ and the correlation coefficient for overseas outsourcing (ρ) and industrial espionage (ϵ) was $r_{oe} = 0.533$.

The test for correlation between outsourcing and espionage is thus

$$t_{oe[8]} = 0.533 / [\sqrt{(1 - 0.533^2)/8}] = 5.034$$

and the calculation of the two-tailed probability that the null hypothesis is true is

$$=T.DIST.2T(t_{oe}, 8) = 0.001***$$

which is extremely significant. We can reasonably reject the null hypothesis; the positive correlation between overseas outsourcing and industrial espionage appears to be real.¹⁰⁰

The negative correlations (outsourcing and profitability; espionage and profitability) need to be converted using the absolute value function (**ABS**). The t-tests are

$$t_{op[8]} = -0.627 / [\sqrt{(1 - 0.627^2)/8}] = -5.927$$

and the calculation of the two-tailed probability that the null hypothesis is true is

$$=T.DIST.2T(ABS(t_{op}), 8) = 0.0004***$$

so the negative correlation between outsourcing and profitability is extremely significant too.

Finally, the correlation between industrial espionage and profitability, -0.761, has a $t_{ep} = -0.761$ with $p(H_0) = 0.0001***$. So that correlation is extremely significant, too.

⁹⁹ The Greek symbol is *rho* which corresponds to our letter *r*. See §7.3 on page 7-3.

¹⁰⁰ Remember this is a completely made-up example! The example does *not* speak to the issue of outsourcing and espionage or anything else in the real world.

9.9 Coefficient of Determination, r^2

A valuable aspect of the correlation coefficient r is that its square, r^2 , known as the *coefficient of determination*, tells us *what proportion of the variation* in one of the variables can be *explained* by the other variable. For example,

- If we learn that the correlation coefficient between the incidence of a type of hacker attack on a network and the occurrence of disk errors on the network disk drives is $r = 0.9$, then $r^2 = 0.81$ and we can assert that in this study, 81% of the variation in one of the variables may be *explained* or *accounted for* by variations in the other. More frequent hacker attacks are positively associated with damaged disk drives; damaged disk drives are associated with a higher frequency of hacker attacks. The 81% figure based on r^2 implies that if we were to define one of the variables as an *independent variable* and the other as a *dependent variable*, we could predict the dependent variable with about 81% of the total variance explained by knowing the value of the dependent variable and 19% left unexplained.
- In our made-up example about outsourcing, espionage and profitability (§9.7), the values of the correlation coefficients can easily be squared in EXCEL to show the coefficients of determination:

Figure 9-22. Coefficients of determination.

COEFFICIENTS OF DETERMINATION	Percentage Overseas Outsourcing	Frequency of Industrial Espionage/1000 employees	Profitability over Last 10 Years
Percentage Overseas Outsourcing	1		
Frequency of Industrial Espionage/1000 employees	28.4%	1	
Profitability over Last 10 Years	39.3%	57.9%	1

Too often, you will hear a journalist or some other statistically naïve person asserting that “the correlation between A and B was 60%, which implies a strong relationship between A and B.” Well, not really: $r = 0.6$ means $r^2 = 0.36$ or in other words, that only 36% of the variation in A can be explained by knowing the value of B or vice versa. All the rest of the variation is unexplained variance. Always mentally square correlation coefficients to estimate the coefficient of determination when you are told about correlations.

Two factors may be positively or negatively correlated because they are *both* determined to some extent by a third, unmeasured factor. Keep in mind is that *correlation does not imply or prove causation* in either direction. For example,

- Just because weight is correlated with age does not mean that weight determines age – or, for that matter, that age determines weight.
- In the outsourcing/espionage/profitability model, there is no implication of causality one way or another.
- And although studies may find a positive correlation between playing violent video games and having aggressive thoughts, the correlation does not mean that playing violent games necessarily causes the increase in aggressivity or that increased aggressivity causes an increase in playing violent video games.¹⁰¹

¹⁰¹ (Anderson and Dill 2000)

9.10 Linear Regression in EXCEL

Sometimes one of the variables in a two-variable data set has been deliberately chosen (the *independent variable*) and the other varies without imposed controls (the *dependent variable*). For example, Figure 9-23 shows the results of an study of the amount of money spent in a month on Internet advertising at Urgonian Corporation and the corresponding monthly sales of the product advertised in the year 2125.

The sales figures are not usually directly under our control (assuming that we aren't selling out our entire production every week) but the amount we spend on advertising is under our control (assuming our marketing manager is not using a Ouija board to determine the spending). This situation is a classic *Model I regression* case in which the X variable – the independent variable – will be the advertising budget and the Y value – the dependent variable – will be the sales figures.

In graphing these data, one can choose the **Insert | Scatter** option:

Figure 9-24. Creating a scatterplot.

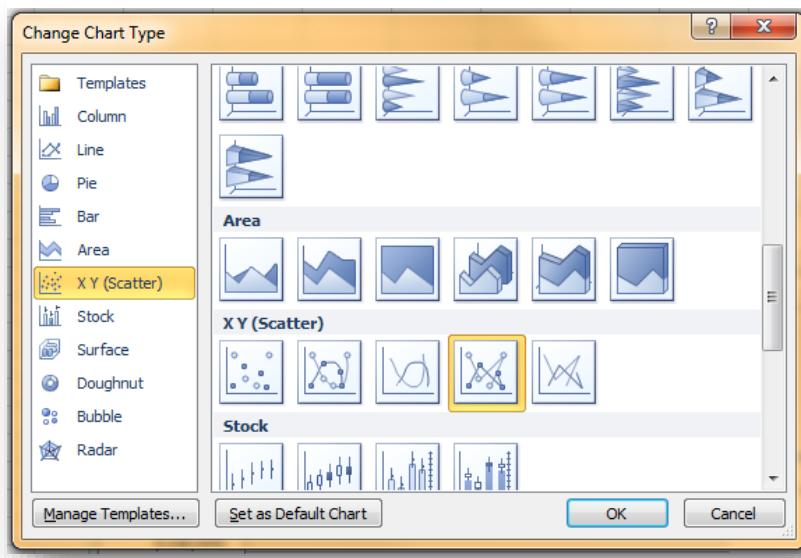
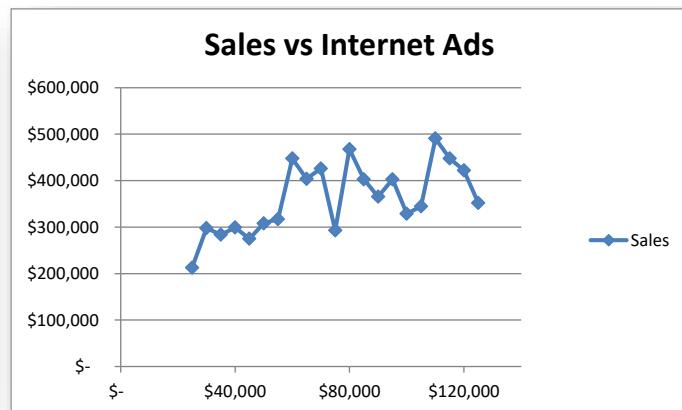


Figure 9-23. Internet ads and sales at Urgonian Corporation in 2125.

Internet Ads	Sales
\$ 25,000	\$ 212,579
\$ 30,000	\$ 297,913
\$ 35,000	\$ 283,758
\$ 40,000	\$ 299,177
\$ 45,000	\$ 274,926
\$ 50,000	\$ 308,150
\$ 55,000	\$ 317,186
\$ 60,000	\$ 447,698
\$ 65,000	\$ 403,792
\$ 70,000	\$ 426,200
\$ 75,000	\$ 292,728
\$ 80,000	\$ 467,736
\$ 85,000	\$ 402,843
\$ 90,000	\$ 365,338
\$ 95,000	\$ 402,883
\$ 100,000	\$ 328,775
\$ 105,000	\$ 344,591
\$ 110,000	\$ 490,599
\$ 115,000	\$ 447,885
\$ 120,000	\$ 422,023
\$ 125,000	\$ 351,895

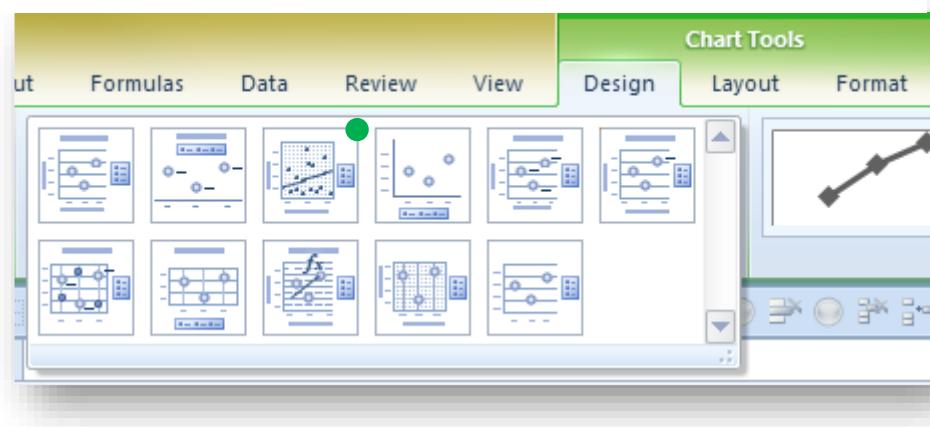
This selection generates a simple XY chart (Figure 9-26) which we can then modify to include a regression line and regression equation:

Figure 9-25. Simple regression without line..



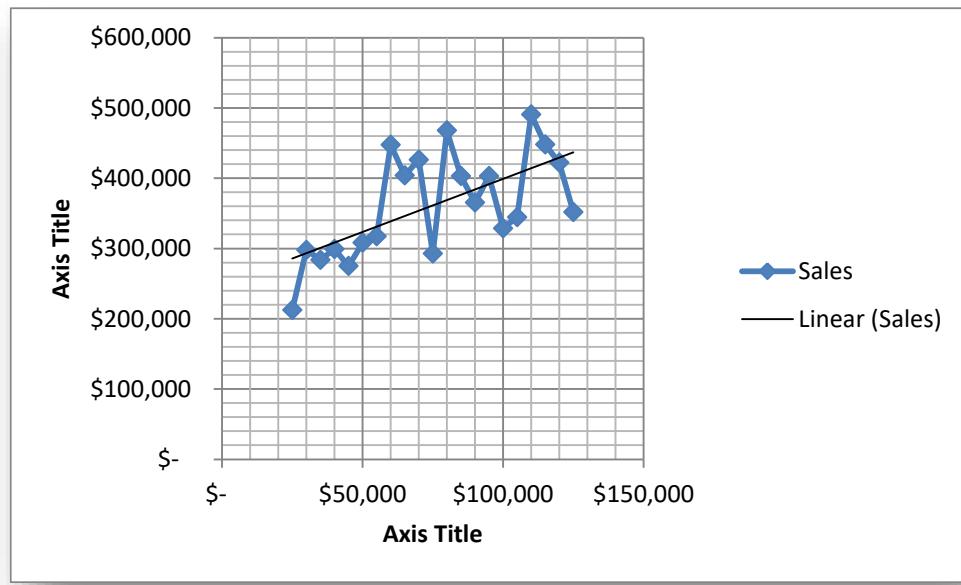
Click on the chart and then use **Chart Tools | Design** to select a version of the chart that shows a linear regression (top row, highlighted with green dot in Figure 9-27):¹⁰²

Figure 9-27. Choosing Layout 3 to show regression line in existing graph.



Instantly, the graph is converted to the form shown in Figure 9-28:

Figure 9-28. Conversion to Type 3 XY plot showing regression line added to raw data.



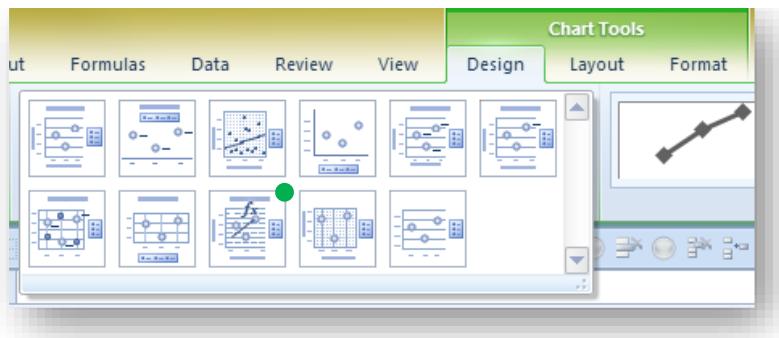
The chart needs additional work, such as labeling the axes and adding a title, but it's a quick and easy way to show the linear regression line without additional computation.

But what if we want to see the regression equation? We have another option.

¹⁰² The green dot does not appear in Excel; it was added in creating the figure.

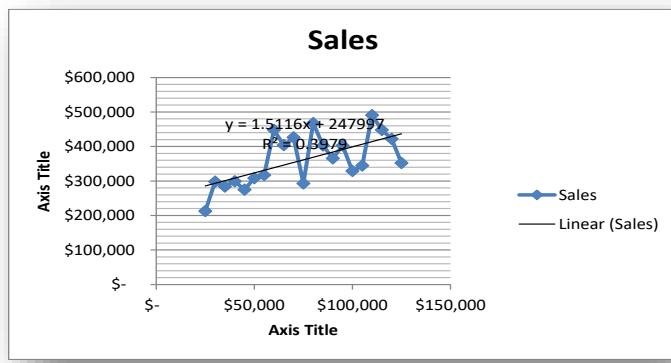
Using Layout 9, shown with the added green dot in Figure 9-29, we can instantly generate a graph that includes the regression equation and the coefficient of determination for the data:

Figure 9-29. Choosing regression line with equation for graph.



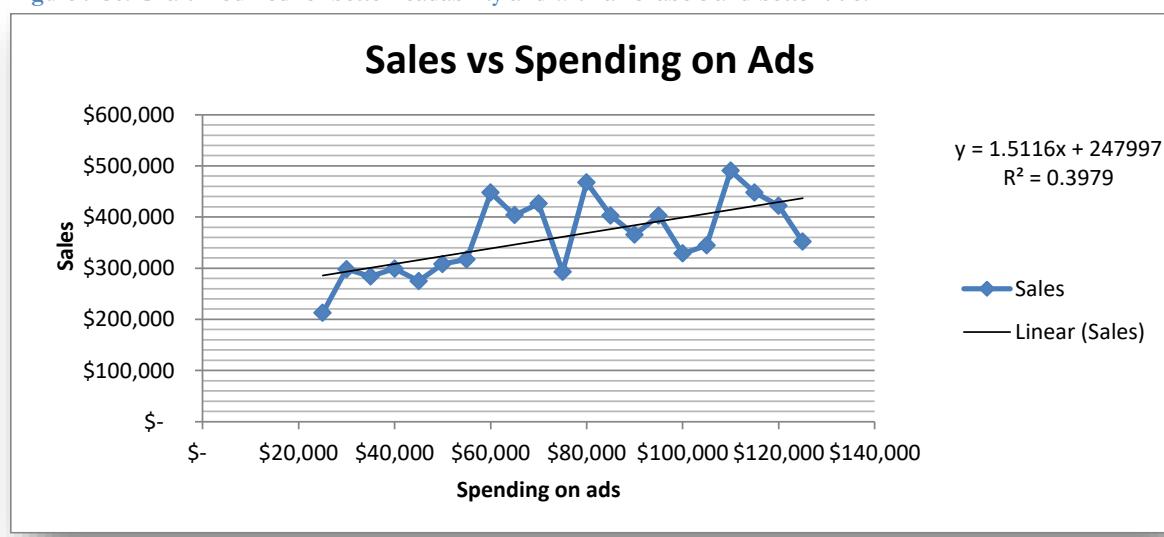
The results (Figure 9-31) are a start:

Figure 9-31. Initial chart produced using Layout 9.



Moving the equation to a more readable area of the chart, stretching the chart sideways, and adding or improving titles, we end up with a presentable representation (Figure 9-30) that includes the regression equation and the coefficient of determination ("R²"):

Figure 9-30. Chart modified for better readability and with axis labels and better title.



9.11 ANOVA with Linear Regression

Typically, we represent the *best-fit linear regression model* as

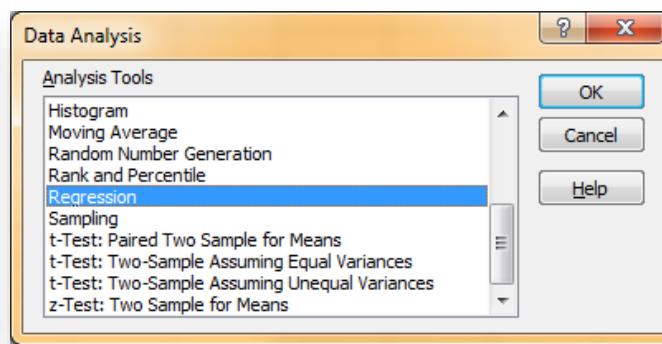
$$\hat{Y} = a + bX + \varepsilon$$

where

- \hat{Y} is the estimated value of the dependent variable for a given independent value X;
- a is the Y-intercept, or the value of Y for $X = 0$;
- b is the regression coefficient, or the amount Y rises for a unit increment in X;
- ε is the residual error, also called the unexplained error – a measure of the average (squared) difference between the predicted values and the observed values.

Figure 9-32 shows the pop-up panel for Regression in the Data Analysis section of EXCEL 2010.

Figure 9-32. Selecting the Regression tool in Data Analysis



For this demonstration, the Regression menu includes the following settings:

Figure 9-33. Regression menu settings for demonstration.

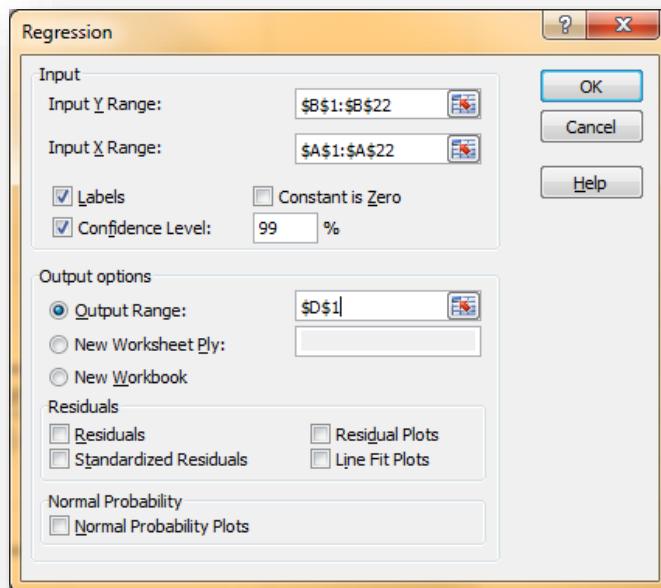


Figure 9-34 shows the results of the **Regression** function, including the *ANOVA with linear regression* table (labeled simply ANOVA in the figure).

Figure 9-34. ANOVA with linear regression for Sales vs Advertising data.

D	E	F	G	H	I	J	K	L	
1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	0.631							
5	R Square	0.398							
6	Adjusted R Square	0.366							
7	Standard Error	59190.873							
8	Observations	21							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	1	43983796124	43983796124	12.554	2.17E-03			
13	Residual	19	66567630337	3503559491					
14	Total	20	110551426461						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
17	Intercept	247997	34505.124	7.187	7.92E-07	175777	320217	149280	346714
18	Internet Ads	1.512	0.427	3.543	2.17E-03	0.619	2.405	0.291	2.732

The regression equation coefficients are

$a = 247,997$ (the Intercept shown in the bottom table in cell E17) and

$b = 1.51$ (the coefficient called “Internet Ads” positioned in cell E22) so the predictive equation (without the error term) is

$$\hat{Y} = 247,997 + 1.51X$$

In the **Regression Statistics** section of the output in Figure 9-34, the key statistic of interest in this course is **R Square** – the coefficient of determination, r^2 (\$0) – which is the variability explained by the regression (**Regress SS**) as a proportion of the total variability (**Total SS**): 0.398 or 39.8%.

The ANOVA section shows us that $F_{[1,19]} = MS_{\text{regression}} / MS_{\text{residual}} = 12.554^{**}$. The regression is highly significant ($p(H_0: \rho=0) = 0.00217$) at the 0.01 level of significance.

The last section of the output includes **Coefficients**: the Y-intercept (a) = 247,997. That represents the predicted sales with zero Internet ads. The 95% confidence limits for a are 175,777 and 320,217. Because the menu in Figure 9-33 includes specification of 99% confidence limits, those are also provided (149,280 and 346,714). These values represent the estimate of how much in sales would occur with zero Internet ads.

The slope (b) is 1.512 (the **Coefficient** in the second row, listing the statistics for **Internet Ads**). The P-value is exactly what is shown in cell I12 of the ANOVA table (2.17E-03) and the confidence limits for b are also shown: 0.619 and 2.405 for the 95% confidence limits; 0.291 and 2.732 for the 99% confidence limits. These values represent the change in expected sales as a proportion of expenditures in Internet ads. Thus at the point estimate $b = 1.512$, the model predicts that every expenditure of a dollar in Internet ads will generate sales of \$1.512 or 151.2% return on investment. On the other hand, the confidence limits also warn that the uncertainty left in the regression ($r^2 = 39.8\%$) means that it is also possible that the return on investment (slope) could be as low as 0.291 or 29.1%. This result indicates that there is a 99% probability that we are correct in asserting that the return on investment in Internet ads will meet or exceed 29.1%.

9.12 Predicted Values in Linear Regression & Confidence Limits

It's easy to generate the predicted values of Y for given values of X by plugging the X-values into the best-fit linear regression equation. Figure 9-35 shows the predicted values of sales for the original list of Internet Ad expenditures (Figure 9-23) using Y-intercept a (\$247,997) and slope b (1.512) calculated by the Data Analysis | Regression tool.

Suppose we want to estimate the sales predicted for expenditures of \$150,000 on Internet ads. We calculate

$$\hat{Y} = \$247,997 + 1.512 * \$150,000 = \$474,734$$

A more involved calculation is to compute the upper and lower $(1 - \alpha)$ confidence limits for a predicted Y (\hat{Y}) for a given X (symbolized by X_i). This measure of uncertainty is smallest at the center of the regression, where the selected X_i is the mean, \bar{X} . As X_i moves further away from the mean, the uncertainty of the prediction increases.

The standard error of \hat{Y} is symbolized $s_{\hat{Y}}$ and is a function of the given value of X_i . It is defined as follows:

$$s_{\hat{Y}} = \sqrt{MS_{\text{residual}} \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{ns_x^2} \right]}$$

- MS_{residual} is the error mean square from the ANOVA;
- n is the sample size;
- X_i is the specific value of the independent variable, X for which we want to compute the predicted value \hat{Y} and its confidence limits;
- \bar{X} is the mean of X;
- s_x^2 is the variance of the values of X in the dataset; can be calculated using =VAR.P(range); unusually, we are using VAR.P instead of VAR.S because it provides the value needed in the computation.

Our example has the following values for these components in the ANOVA with regression for our example:

- $MS_{\text{residual}} = 3,503,559,491$
- $n = 21$
- $X_i = 150,000$
- $\bar{X} = 75,000$
- $s_x^2 = 916,666,667$

The calculation yields $s_{\hat{Y}} = 34,505$.

The distribution of \hat{Y} follows Student's-t with $v = n - 2$ degrees of freedom.

Figure 9-35. Predicted sales as function of Internet ad expenditures.

Internet Ads	Pred. Sales
\$ 25,000	\$285,787
\$ 30,000	\$293,344
\$ 35,000	\$300,902
\$ 40,000	\$308,460
\$ 45,000	\$316,018
\$ 50,000	\$323,576
\$ 55,000	\$331,134
\$ 60,000	\$338,692
\$ 65,000	\$346,250
\$ 70,000	\$353,808
\$ 75,000	\$361,365
\$ 80,000	\$368,923
\$ 85,000	\$376,481
\$ 90,000	\$384,039
\$ 95,000	\$391,597
\$ 100,000	\$399,155
\$ 105,000	\$406,713
\$ 110,000	\$414,271
\$ 115,000	\$421,829
\$ 120,000	\$429,387
\$ 125,000	\$436,944

Computation of the confidence limits for \hat{Y} can use the =CONFIDENCE.T(alpha, standard_dev, size) function from EXCEL 2010. The only wrinkle is that =CONFIDENCE.T is defined for computing confidence limits of the *mean*, and it therefore assumes that the degrees of freedom are *size* – 1. Because the degrees of freedom of $s_{\hat{Y}}$ are $v = n - 2$, we have to trick the function by entering the *size* parameter as $n - 1$ to force the function to use $v = n - 2$ for its calculation.

The =CONFIDENCE.T value is one-half the confidence interval. Thus lower (L_1) and upper (L_2) $(1 - \alpha)$ confidence limits are

$$L_1 = \hat{Y} - \text{CONFIDENCE.T}(\alpha, s_{\hat{Y}}, n-1)$$

$$L_2 = \hat{Y} + \text{CONFIDENCE.T}(\alpha, s_{\hat{Y}}, n-1)$$

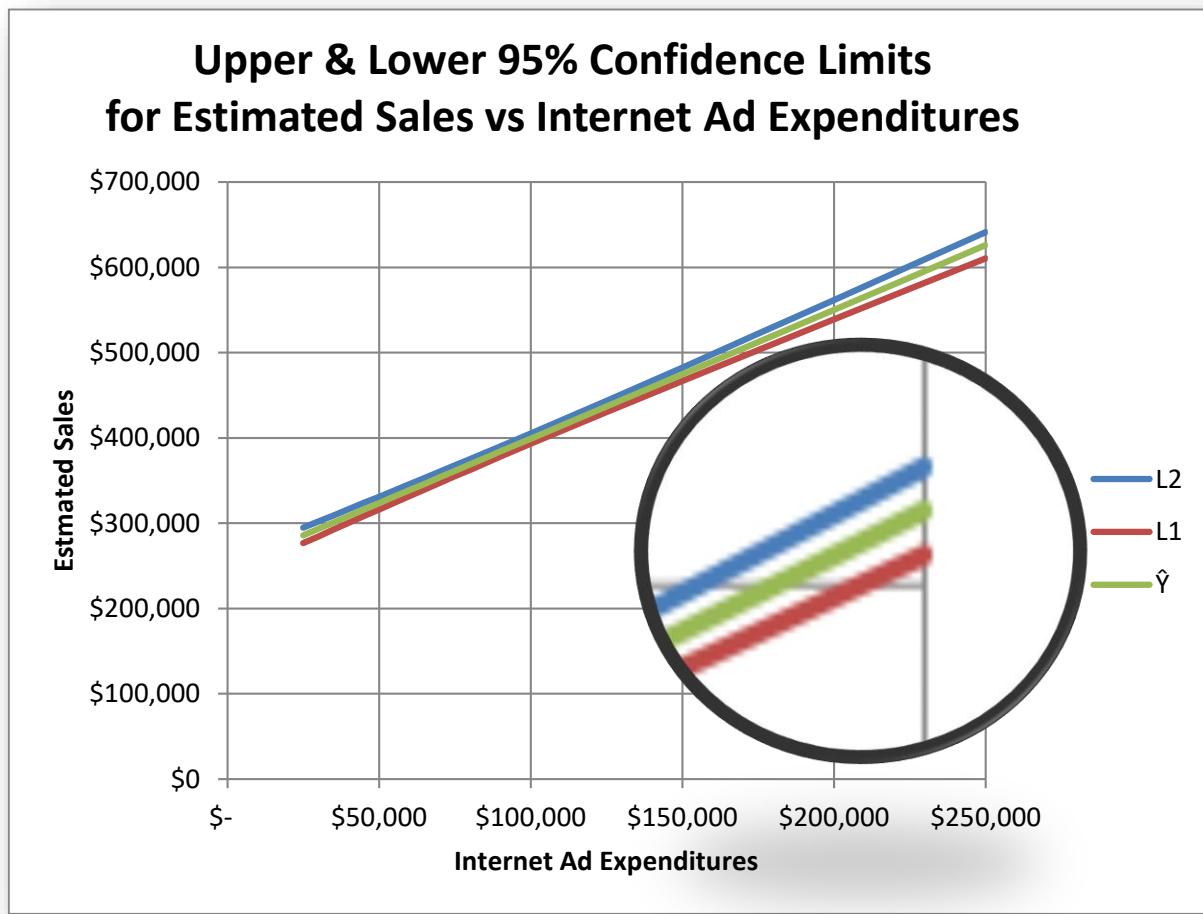
In our example, the 95% confidence limits for the estimated value $\hat{Y} = \$474,734$ for $X_i = \$150,000$ are

$$L_1 = \$458,585$$

$$L_2 = \$490,883$$

After calculating the upper and lower 95% confidence limits for all the values in the original data chart, it's easy to create a chart illustrating the bowed (parabolic) nature of the upper and lower confidence limits. Figure 9-36 shows the modest widening of the confidence limits around the estimated sales. The values include Internet Ad expenditures reaching up to \$250,000 to demonstrate the curve of the confidence limits. The circular inset expands the right-hand side of the predictions to illustrate the divergence between upper bound (blue) expected value (green) and lower bound (red).

Figure 9-36. Confidence limits get wider away from the mean.



10 Analyzing Frequency Data

Often we have to evaluate whether counted data – frequency data – fit a predictive model. In other words, we must test the hypothesis that any deviations from expectations are due to chance alone – sampling variability that accidentally allows the tallies in each category to vary from a perfect match with theory.

10.1 Computing Expected Frequencies for Statistical Distributions

One common question about frequency data is whether they fit a preconceived theoretical distribution – say, the Normal distribution or a Student's-t distribution. The source of predicted or expected values may be the results of previous research, calculations based on theoretical models, and calculations based on theoretical statistical distribution functions. For example, one might want to know if a certain data set were distributed in accordance with the Normal distribution. We can set up a number of categories which have equal percentages of the theoretical population allocated to them by using the =NORM.INV(p, μ, σ) function in EXCEL, and supplying the parametric mean and standard deviation of the desired distribution for each portion of the Normal curve to the left of the critical values. For example, suppose we are testing the fit of an observed distribution to a theoretical Normal distribution with parametric mean $\mu = 50$ and parametric standard deviation $\sigma = 10$. Figure 10-1 shows the computations.

With these class boundaries in place, we can now count the number of observations that fit into these classes

Figure 10-1. Computing class limits for a Normal frequency distribution using Excel.

Normal Curve Limits		Mean:	50
		SDEV:	10
% to left of boundary		Class upper boundary	
10%		37.18	=NORM.INV(A4,\$E\$1,\$E\$2)
20%		41.58	
30%		44.76	
40%		47.47	
50%		50.00	
60%		52.53	
70%		55.24	
80%		58.42	
90%		62.82	

and pursue the investigation of the degree to which the observed frequencies match the predicted (theoretical) frequencies. Figure 10-2 shows the results of an empirical study. How do we tell if what we observed fits the theoretical distribution? We use a goodness-of-fit test.

Figure 10-2. Observed proportions for test of goodness of fit to Normal curve.

Class upper boundary	Expected %	Observed %
37.18	10%	13%
41.58	10%	10%
44.76	10%	14%
47.47	10%	14%
50.00	10%	13%
52.53	10%	13%
55.24	10%	9%
58.42	10%	6%
62.82	10%	8%

Another common question is whether the observed *frequencies* of *attributes* for different *groups* are similar or different; e.g., are the frequencies of opinions at five different levels on a Likert scale (§1.8) the same in different groups of subjects defined by demographic criteria? Figure 10-3 shows the results of a survey of responses for people classified by their degree of educational attainment.

Figure 10-3. Frequency distributions of responses on Likert scale for groups with different educational attainment.

Educational Level	Responses on Likert Scale					Total
	-2	-1	0	1	2	
< High School	153	19	9	5	3	189
HSL or GDE	145	41	20	10	6	222
Associates	52	70	35	8	8	173
Baccalaureate	2	28	42	35	32	139
Masters	0	8	17	34	26	85
Doctorate	0	0	7	22	20	49
Post-doctoral	0	0	2	8	11	21

Simply looking at the data, it would seem that there *may* be a shift towards the positive side of the Likert scale with rising educational level – but how do we quantify the likelihood that these observations are simply due to random sampling with no parametric difference of responses among the groups? Such an analysis is called a *test of independence* which also uses the concept of *goodness of fit*.

10.2 The Chi-Square Goodness-of-Fit Test

One of the questions of interest to financial analysts looking at trading patterns within the Solar System is the extent to which global trading conforms to a formal model they have developed to predict the frequency of what are called spiral disasters in which computerized trading programs enter positive-feedback loops despite programming designed to prevent such catastrophic events. A research group has collected 24 months of data and has counted how many spiral disasters occurred in each month. They then computed the predicated frequencies according to their model and now want to know if the observed data are consistent with their model. The three columns on the left of Figure 10-4 show their data, including the expected frequencies they computed.

Notice that the sum of the expected frequencies *must* equal the sum of the observed frequencies; calculating these sums provides a way of checking one's computations to be sure that everything is in order.

The fundamental operation in the *chi-square goodness-of-fit test* is to compute the square of the *observed frequencies minus the expected frequencies* and divide each difference by the expected frequency and add up all these quotients:

$$X_v^2 = \sum \frac{(o - e)^2}{e}$$

If the observations are random samples from a population that follows the expected frequencies, then X^2 is distributed as $\chi^2_{[v]}$, a chi-square distribution with $v = (r - 1)(c - 1)$ degrees of freedom where r is the number of rows (24 here) and c is the number of columns (2 in this case). Thus in this case $v = 23$.

The hypotheses under test in this method are

H0: the observed frequencies fit the expected distribution

H1: the observed frequencies do not fit the expected distribution.

In EXCEL, we compute $P\{H_0\} =$

$$=CHISQ.DIST.RT(35.3,21) = 0.0262^*$$

There is a statistically significant deviation of the observed frequency distribution from expectation at the 0.05 level of significance. The model used for predicting the number of spiral disasters does not appear to be a good fit to the observed data. Back to the drawing hologram for the analysts!

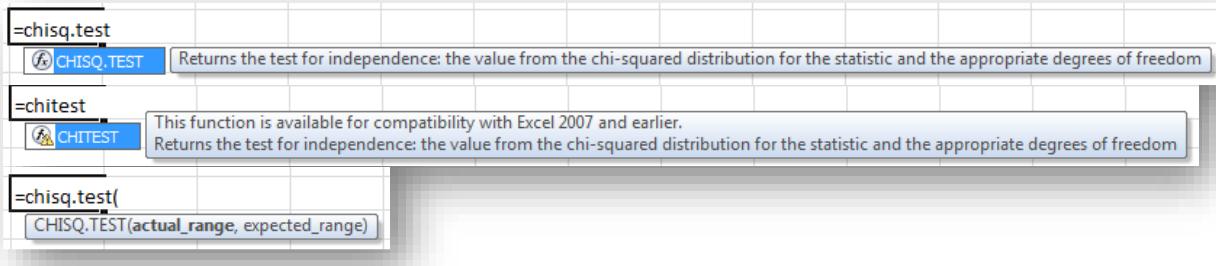
Figure 10-4. Data for goodness-of-fit test.

Spiral Disasters Analysis			
Month	Observed Spirals	Expected Spirals	$(o-e)^2/e$
1	2	3.665625	0.756844
2	4	4.096875	0.002291
3	4	4.528125	0.061596
4	6	4.959375	0.218354
5	7	5.390625	0.48048
6	2	5.821875	2.508939
7	2	6.253125	2.892805
8	3	6.684375	2.030799
9	1	7.115625	5.256161
10	2	7.546875	4.076896
11	6	7.978125	0.490463
12	4	8.409375	2.312013
13	16	8.840625	5.797854
14	12	9.271875	0.802714
15	11	9.703125	0.173334
16	11	10.134375	0.079387
17	10	10.565625	0.03028
18	10	10.996875	0.090367
19	13	11.428125	0.216203
20	11	11.859375	0.062274
21	17	12.290625	1.804482
22	19	12.721875	3.098195
23	16	13.153125	0.61618
24	18	13.584375	1.435307
SUMS:	207	207.0	35.288769
Count:	24	=a	

10.3 The EXCEL =CHISQ.TEST Function

EXCEL 2010 includes functions that perform a chi-square test for goodness of fit on two arrays, the observed frequencies and the expected frequencies (Figure 10-5):

Figure 10-5. Excel 2010 chi-square goodness-of-fit functions.



The =CHISQ.TEST function returns the probability that the null hypothesis **H0: no deviation from expectation** is true. It does so by computing the sum $X_v^2 = \sum \frac{(o-e)^2}{e}$ with $v = (r-1)(c-1)$ degrees of freedom, as discussed in §10.2.

The example in Figure 10-6 shows the results of tests of tensile strength on 400 structural rods used in a starship; the distribution of observed tensile strengths shows $P(H_0) = 0.958$ using the =CHISQ.TEST function to compute the probability directly (row 8). There is no reason to suspect that the rods deviate from the standard.

If the value of X_v^2 is required, one can use the =CHISQ.INV function with v degrees of freedom to compute it.

For interest, row 9 shows the calculation of X_v^2 directly and then row 10 shows how the $P(H_0)$ is derived using the =CHISQ.DIST.RT function. The results are identical.

Figure 10-6. Tests of goodness of fit for tensile strength of structural rods.

	A	B	C	D	E
1	Tensile Strength (kg/mm)	Expected %	Expected Frequency	Observed Frequency	$((o-e)^2)/e$
2	80	9.096%	36.385	37	0.010
3	100	40.879%	163.516	165	0.013
4	120	40.879%	163.516	159	0.125
5	140	9.146%	36.583	39	0.160
6	Total		400.000	400.000	0.308
7					
8	P(H0) from func:	0.958	=CHISQ.TEST(D2:D7,C2:C7)		
9	χ^2 from P(H0):	0.308	=CHISQ.INV.RT(B8,3)		
10	χ^2 from calc:	0.308	=E6		
11	P(H0) from calc:	0.958	=CHISQ.DIST.RT(E6,3)		

10.4 Two-Way Tests of Independence Using Chi-Square

One of the most common expectations we use in statistics is that two factors are independent; that is, that the level of one variable has no effect on the level of another variable. When we use categorical data, such as counts of how many people respond to questions using a Likert scale of preferences, we can test for independence of their responses from other factors using chi-square tests of independence.

In §4.11, we discussed the results of a customer-satisfaction survey by the Urgonian Corporation. Concerned by the results, which suggested high negative responses to their products in some areas of the Solar System, the Corporation runs an experiment a year later to see if the regional differences they noted are still in effect. The results are shown in Figure 10-7.

To test

Figure 10-7. Observed frequencies of responses in different sectors.

	A	B	C	D	E
1	SURVEYS 2219.05.20	Positive	Negative	Total Ret'd	% Positive
2	HQ (Beijing, Terra)	231	77	308	75.0%
3	Mare Imbrium, Luna	224	67	291	77.0%
4	Tycho Colony, Luna	249	93	342	72.8%
5	Olympus Mons, Mars	222	63	285	77.9%
6	Southern Ice Cap, Mars	269	127	396	67.9%
7	Callisto Colony, Jupiter	88	39	127	69.3%
8	Europa Colony, Jupiter	158	50	208	76.0%
9	Totals	1,441	516	1,957	73.6%

hypothesis of independence between the location of the respondents and their responses, we need to compute the expected frequencies for the positive and negative tallies. These expected values are shown in Figure 10-8.

Figure 10-8. Expected frequencies of responses in different sectors under the hypothesis of independence.

Expected Frequencies	Positive	Negative	Check Totals	
HQ (Beijing, Terra)	226.79	81.21	308.00	OK
Mare Imbrium, Luna	214.27	76.73	291.00	OK
Tycho Colony, Luna	251.83	90.17	342.00	OK
Olympus Mons, Mars	209.85	75.15	285.00	OK
Southern Ice Cap, Mars	291.59	104.41	396.00	OK
Callisto Colony, Jupiter	93.51	33.49	127.00	OK
Europa Colony, Jupiter	153.16	54.84	208.00	OK
Check Totals	1,441.00	516.00	1,957.00	OK

The expected frequency of positive responses in all locations would be the observed frequency of positive results in the total survey (73.6%), calculated by dividing the total positive results (1,441 in **cell B9**) by the total responses (1,957 in **cell D9**).

Therefore the expected number of positive responses in a specific location would be 73.6% of that location's total response; e.g., the HQ responses, totaling 308 (**cell D2**), would be expected to show $0.736 \times 308 = 226.79$ positive responses. Because this is a theoretical number, the impossible decimal fractions don't cause a problem – they are simply part of the calculations. To be sure that our calculations are correct, it is advisable to include check totals and to verify that they match the original marginal totals.

Figure 10-9 shows the EXCEL formulas used for the calculations. In defining the first expected frequency, in **cell B2**, notice that the marginal total for the number of positive responses in **cell B9** is locked by row using **B\$9** in the computation. Similarly, the marginal total for the total number of responses from Beijing (**cell D2**) is locked by column (**\$D2**). The reference to the total number of responses in **cell D9** is locked by both row and column (**\$D\$9**).

Figure 10-9. Formulas for expected frequencies displayed.

	A	B	C	D	E
1	SURVEYS 2219.05.20	Positive	Negative	Total Ret'd	% Positive
2	HQ (Beijing, Terra)	231	77	=SUM(B2:C2)	=+B2/D2
3	Mare Imbrium, Luna	224	67	=SUM(B3:C3)	=+B3/D3
4	Tycho Colony, Luna	249	93	=SUM(B4:C4)	=+B4/D4
5	Olympus Mons, Mars	222	63	=SUM(B5:C5)	=+B5/D5
6	Southern Ice Cap, Mars	269	127	=SUM(B6:C6)	=+B6/D6
7	Callisto Colony, Jupiter	88	39	=SUM(B7:C7)	=+B7/D7
8	Europa Colony, Jupiter	158	50	=SUM(B8:C8)	=+B8/D8
9	Totals	=SUM(B2:B8)	=SUM(C2:C8)	=SUM(D2:D8)	=+B9/D9
10					
11					
12	Expected Frequencies	Positive	Negative	Check Totals	
13	HQ (Beijing, Terra)	=B\$9*\$D2/\$D\$9	=C\$9*\$D2/\$D\$9	=SUM(B13:C13)	=IF(D13=D2,"OK","ERROR")
14	Mare Imbrium, Luna	=B\$9*\$D3/\$D\$9	=C\$9*\$D3/\$D\$9	=SUM(B14:C14)	=IF(D14=D3,"OK","ERROR")
15	Tycho Colony, Luna	=B\$9*\$D4/\$D\$9	=C\$9*\$D4/\$D\$9	=SUM(B15:C15)	=IF(D15=D4,"OK","ERROR")
16	Olympus Mons, Mars	=B\$9*\$D5/\$D\$9	=C\$9*\$D5/\$D\$9	=SUM(B16:C16)	=IF(D16=D5,"OK","ERROR")
17	Southern Ice Cap, Mars	=B\$9*\$D6/\$D\$9	=C\$9*\$D6/\$D\$9	=SUM(B17:C17)	=IF(D17=D6,"OK","ERROR")
18	Callisto Colony, Jupiter	=B\$9*\$D7/\$D\$9	=C\$9*\$D7/\$D\$9	=SUM(B18:C18)	=IF(D18=D7,"OK","ERROR")
19	Europa Colony, Jupiter	=B\$9*\$D8/\$D\$9	=C\$9*\$D8/\$D\$9	=SUM(B19:C19)	=IF(D19=D8,"OK","ERROR")
20	Check Totals	=SUM(B13:B19)	=SUM(C13:C19)	=SUM(D13:D19)	=IF(D20=D9,"OK","ERROR")
21		=IF(B20=B9,"OK","ERROR")	=IF(C20=C9,"OK","ERROR")		

When the formula in **cell B2** is propagated to **cell C2** for the expected number of negative responses in Beijing, the locked components stay the same but the pointers shift appropriately; thus the calculation uses the number of total number of negative responses (**cell C9**) instead of the original **cell B9** corresponding to total positive responses.

When we propagate both **cells B2** and **C2** downward, the locked components ensure that the correct rows are used; e.g., for **Mare Imbrium** (**row 3**), the calculations of expected frequencies refer to **cell D3** for the total number of responses from that location.

Through these means, we need only to write the formula for the first cell (**B2**) and then propagate horizontally, then propagate the first row downwards. This process minimizes the errors that could occur by manually writing out formula after formula for all the cells – and shortens the process to a few seconds.

The degrees of freedom for a two-way test of independence are the product of the degrees of freedom of the number of rows and of the number of columns: $\nu = (r-1)(c-1)$. In our case, we have 7 rows and 2 columns, so $\nu = 6*1 = 6$.

Using the EXCEL statistical function =CHISQ.TEST(array1, array2), we identify **array1** as the matrix of observed results and **array2** as the matrix of expected results. The formula instantly computes the probability of obtaining the calculated value (or higher) of the chi-square test of independence (which it does not reveal) by chance alone if there is no relationship between the two variables (in our case, location and response pattern). To determine the actual value of the chi-square computation, we can use the =CHISQ.INV.RT(p, df) function in EXCEL, which gives us the value of the chi-square distribution with *df* degrees of freedom for which a $P\{H_0\}$ of *p* has been obtained.

Figure 10-10 shows the results of the calculation in EXCEL 2010 and includes the formulas displayed.

Based on this test of independence, the Urgonian Corporation statisticians find a significant probability

Figure 10-10. Excel 2010 calculations for test of independence using chi-square.

P(H0):	0.0399	=CHISQ.TEST(B2:C8,B13:C19)		
X ² :	13.207	=CHISQ.INV.RT(C23,(COUNTA(A2:A8)-1)*(COUNTA(B1:C1)-1))		

($P\{H_0\} = 0.0399^*$) at the 0.05 level of significance that the null hypothesis is false. There appears to be a statistically significant effect of location on the proportion of positive responses in the survey data.

In summary, the chi-square test of independence requires us to

- (1) Lay out an array of observed results;
- (2) Compute expected results based on the proportions defined by the marginal totals;
- (3) Compute the degrees of freedom for the chi-square test;
- (4) Use the =CHISQ.TEST function on the observed array and the expected array to compute the $P\{H_0\}$;
- (5) Use the =CHISQ.INV.RT function to find the value of the chi-square statistic.
- (6) Calculate $\nu = (\text{rows}-1)(\text{columns}-1)$.

As a matter of interest, Figure 10-11 shows the detailed calculation of the chi-square statistic for these data by summing the expression $((o-e)^2)/e$ for all cells. As you can see, the lower-right corner cell has exactly the same result (13.207) as the calculation of X² shown in Figure 10-10

Figure 10-11. Calculating the chi-square for independence the long way.

(o-e) ² /e	Positive	Negative	Totals
HQ (Beijing, Terra)	0.0782	0.2183	0.296
Mare Imbrium, Luna	0.4416	1.2333	1.675
Tycho Colony, Luna	0.0317	0.0885	0.120
Olympus Mons, Mars	0.7029	1.9631	2.666
Southern Ice Cap, Mars	1.7497	4.8862	6.636
Callisto Colony, Jupiter	0.3251	0.9080	1.233
Europa Colony, Jupiter	0.1531	0.4277	0.581
Totals	3.482	9.725	13.207

11Introduction to Minitab

12Multifactorial Analysis

12.1 Nested ANOVA

12.2 Two-factor ANOVA and Introduction to Interaction Terms

12.3 Multi-way ANOVA and Analyzing Interactions

12.4 Analyzing Nested and Multidimensional Contingency Tables

12.5 Multiple Linear Regression

13 Assumptions of Parametric Analysis

13.1 Normality & Homoscedasticity Revisited

13.2 Data Transformations

13.2.1 Arcsin Transform for Extreme Percentages

13.2.2 Log Transform for Heteroscedastic Data

13.3 Curvilinear Regression

13.3.1 Using Standard Transforms

13.3.2 Exploratory Transforms

13.4 Non-Parametric Analysis Tools

13.4.1 Kruskal-Wallis for Single-classification Ranked Data

13.4.2 Friedman's Method for Randomized Blocks

13.4.3 Mann-Whitney U-test

13.4.4 Komogorov-Smirnov Two-sample Test

13.4.5 Wilcoxon's Signed Ranks Test

13.4.6 Kendall's Robust Line-Fit Method

13.4.7 Kendall's Coefficient of Rank Correlation

13.4.8 Spearman's Coefficient of Rank Correlation

13.4.9 Kendall's Coefficient of Concordance

14Exploratory Data Analysis

14.1 Moving Averages for Time Series

14.2 Partial Correlation Coefficients

14.3 Multiple Comparisons in ANOVA

14.4 Factor Analysis

14.5 Monte Carlo Simulation

14.6 Meta-Analysis

14.7 *A Posteriori* Hypothesis Testing

15 Experimental Design Fundamentals

15.1 Classic Pitfalls

15.2 Estimating Appropriate Sample Sizes

15.3 Control Groups and Double-Blind Studies

15.4 Survey Design

15.5 Focus Groups

15.6 Computer-Aided Consensus™

15.7 Interviews

15.8 Computer-Aided Thematic Analysis

16Miscellaneous Decision-Support Methods

16.1 Heuristic Weighting Tables

16.2 Operational Matrices

17Concluding Remarks

18Bibliography

19

- Adler, M. J., and C. Van Doren. 1972. *How to Read a Book*. Revised. Touchstone.
- Alexander, M. 2013. "Why Excel has Multiple Quartile Functions and How to Replicate the Quartiles from R and Other Statistical Packages." *Bacon Bits*. DataPig Technologies. 11 15. Accessed 02 14, 2015. <http://datapigtechnologies.com/blog/index.php/why-excel-has-multiple-quartile-functions-and-how-to-replicate-the-quartiles-from-r-and-other-statistical-packages/>.
- Anderson, C. A., and K. E. Dill. 2000. "Video Games and Aggressive Thoughts, Feelings, and Behavior in the Laboratory and in." *Journal of Personality and Social Psychology* 78 (4): 772-790.
- Associated Press. 2011. "Turkish quake reveals shoddy building techniques." *USA TODAY*. 10 26. Accessed 08 10, 2012. <http://www.usatoday.com/news/world/story/2011-10-26/turkey-quake/50921970/1>.
- Bost, J. 2003. "How to insert symbols and special characters into a Word document (video)." *Office in Education*. 12. Accessed 08 06, 2012. <http://blogs.office.com/b/office-education/archive/2010/12/03/how-to-insert-symbols-and-special-characters-into-a-word-document.aspx>.
- Cambell, H. A. 2011. "What is a mole?" *Kiwi Web: Chemistry & New Zealand*. Accessed 07 16, 2012. <http://www.chemistry.co.nz/mole.htm>.
- Carlberg, C. 2011. *Statistical Analysis: Microsoft Excel 2010*. Que.
- CollegeBoard. 2012. "How the SAT Is Scored." *College Board SAT*. Accessed 07 22, 2012. <http://sat.collegeboard.org/scores/how-sat-is-scored>.
- Encyclopaedia Britannica. 2012. "Student's t-test." *Encyclopaedia Britannica Online*. Accessed 08 12, 2012. <http://www.britannica.com/EBchecked/topic/569907/Students-t-test>.
- EuSpRIG – European Spreadsheet Risks Interest Group. 2012. "EuSpRIG Original Horror Stories #073 & #083." *EuSpRIG*. Accessed 07 24, 2012. <http://www.eusprig.org/stories.htm>.
- Fowler, M. 2009. "The Speed of Light." *Galileo and Einstein*. 09 01. Accessed 07 16, 2012. <http://galileo.phys.virginia.edu/classes/109N/lectures/spedlite.html>.
- Gilson, D., and C. Perot. 2011. "It's the Inequality, Stupid: Eleven charts that explain what's wrong with America." *Mother Jones: Plutocracy Now*. 03 & 04. Accessed 08 05, 2012. <http://www.motherjones.com/politics/2011/02/income-inequality-in-america-chart-graph>.
- Gray, J. J. 2012. "Carl Friedrich Gauss." *Encyclopaedia Britannica Online*. Accessed 08 11, 2012. <http://www.britannica.com/EBchecked/topic/227204/Carl-Friedrich-Gauss>.
- Lund, P. 2010. "Letter to the Editor." *Guardian Weekly*, 07 16: 23.
- Morgan, S. L. 2010. "Tutorial on the Use of Significant Figures." *University of South Carolina Faculty Pages | Morgan | Analytical Chemistry*. Accessed 07 16, 2012. <http://www.chem.sc.edu/faculty/morgan/resources/sigfigs/index.html>.
- Nykamp, D. Q. nd. "The idea of a probability density function." *Math Insight*. nd nd. Accessed 03 30, 2016. http://mathinsight.org/probability_density_function_idea.
- O'Connor, J. J., and E. F. Robertson. 2003. "Karl Pearson." *The MacTutor History of Mathematics archive*. 10. Accessed 08 18, 2012. <http://www-history.mcs.st-andrews.ac.uk/Biographies/Pearson.html>.
- Peltier, J. 2011. "Excel Box and Whisker Diagrams (Box Plots)." *Peltier Tech Blog*. 06 07. Accessed 08 08, 2012. <http://peltiertech.com/WordPress/excel-box-and-whisker-diagrams-box-plots/>.

- Rohlf, F. J., and R. R. Sokal. 1981. *Statistical Tables*. Second. New York, NY: W. H. Freeman.
- Sokal, R. R., and F. J. Rohlf. 1981. *Biometry: The Principles and Practice of Statistics in Biological Research*. 2nd. New York: W. H. Freeman & Company.
- . 2012. *Biometry: The Principles and Practice of Statistics in Biological Research*. 4th. New York: W. H. Freeman and Company.
- Stathopoulos, V. 2012. "Space Shuttle Challenger Disaster." *Aerospace Guide*. 06 02. Accessed 07 22, 2012.
http://www.aerospaceguide.net/spaceshuttle/challenger_disaster.html.
- Trochim, William M. K. 2006. "Scaling." *Research Methods Knowledge Base*. 10 20. Accessed 07 12, 2012.
<http://www.socialresearchmethods.net/kb/scaling.php>.
- University of Exeter. 2012. "Theoretical PhysicsPi." *Physics & Astronomy Quantum Systems and Nanomaterials Group*. 07 16. Accessed 07 16, 2012. <http://newton.ex.ac.uk/research/qsystems/collabs/pi/>.
- Washington's Blog. 2008. "Hedge Funds' Derivatives Exposure and Margin Calls Driving Stock Market Crash." *Washington's Blog*. 10 15. Accessed 07 12, 2012.
<http://georgewashington2.blogspot.com/2008/10/hedge-funds-derivatives-exposure.html>.
- Wolfram Mathworld. 2012. "e." *Wolfram Mathworld*. 07 16. Accessed 07 16, 2012.
<http://mathworld.wolfram.com/e.html>.
- . 2012. "Googolplex." *Wolfram MathWorld*. 07 02. Accessed 07 12, 2012.
<http://mathworld.wolfram.com/Googolplex.html>.

