Advance Praise for Head First Statistics

"*Head First Statistics* is by far the most entertaining, attention-catching study guide on the market. By presenting the material in an engaging manner, it provides students with a comfortable way to learn an otherwise cumbersome subject. The explanation of the topics is presented in a manner comprehensible to students of all levels."

- Ariana Anderson, Teaching Fellow/PhD candidate in Statistics, UCLA

"*Head First Statistics* is deceptively friendly. Breeze through the explanations and exercises and you just may find yourself raising the topic of normal vs. Poisson distribution in ordinary social conversation, which I can assure you is not advised!"

- Gary Wolf, Contributing Editor, Wired Magazine

"Dawn Griffiths has split some very complicated concepts into much smaller, less frightening, bits of stuff that real-life people will find very easy to digest. Lots of graphics and photos make the material very approachable, and I have developed quite a crush on the attractive lady model who is asking about gumballs on page 458."

- Bruce Frey, author of Statistics Hacks

"Head First is an intuitive way to understand statistics using simple, real-life examples that make learning fun and natural."

— Michael Prerau, computational neuroscientist and statistics instructor, Boston University

"Thought Head First was just for computer nerds? Try the brain-friendly way with statistics and you'll change your mind. It really works."

- Andy Parker

"This book is a great way for students to learn statistics—it is entertaining, comprehensive, and easy to understand. A perfect solution!"

- Danielle Levitt

"Down with dull statistics books! Even my cat liked this one."

- Cary Collett

Praise for other Head First books

"Kathy and Bert's *Head First Java* transforms the printed page into the closest thing to a GUI you've ever seen. In a wry, hip manner, the authors make learning Java an engaging 'what're they gonna do next?' experience."

-Warren Keuffel, Software Development Magazine

"Beyond the engaging style that drags you forward from know-nothing into exalted Java warrior status, Head First Java covers a huge amount of practical matters that other texts leave as the dreaded "exercise for the reader..." It's clever, wry, hip and practical—there aren't a lot of textbooks that can make that claim and live up to it while also teaching you about object serialization and network launch protocols. "

—Dr. Dan Russell, Director of User Sciences and Experience Research IBM Almaden Research Center (and teaches Artificial Intelligence at Stanford University)

"It's fast, irreverent, fun, and engaging. Be careful-you might actually learn something!"

—Ken Arnold, former Senior Engineer at Sun Microsystems Co-author (with James Gosling, creator of Java), *The Java Programming Language*

"I feel like a thousand pounds of books have just been lifted off of my head."

-Ward Cunningham, inventor of the Wiki and founder of the Hillside Group

"Just the right tone for the geeked-out, casual-cool guru coder in all of us. The right reference for practical development strategies—gets my brain going without having to slog through a bunch of tired stale professor-speak."

-Travis Kalanick, Founder of Scour and Red Swoosh Member of the MIT TR100

"There are books you buy, books you keep, books you keep on your desk, and thanks to O'Reilly and the Head First crew, there is the penultimate category, Head First books. They're the ones that are dog-eared, mangled, and carried everywhere. Head First SQL is at the top of my stack. Heck, even the PDF I have for review is tattered and torn."

- Bill Sawyer, ATG Curriculum Manager, Oracle

"This book's admirable clarity, humor and substantial doses of clever make it the sort of book that helps even non-programmers think well about problem-solving."

- Cory Doctorow, co-editor of Boing Boing Author, Down and Out in the Magic Kingdom and Someone Comes to Town, Someone Leaves Town

Praise for other Head First books

"I received the book yesterday and started to read it...and I couldn't stop. This is definitely très 'cool.' It is fun, but they cover a lot of ground and they are right to the point. I'm really impressed."

— Erich Gamma, IBM Distinguished Engineer, and co-author of *Design Patterns*

"One of the funniest and smartest books on software design I've ever read."

- Aaron LaBerge, VP Technology, ESPN.com

"What used to be a long trial and error learning process has now been reduced neatly into an engaging paperback."

- Mike Davidson, CEO, Newsvine, Inc.

"Elegant design is at the core of every chapter here, each concept conveyed with equal doses of pragmatism and wit."

- Ken Goldstein, Executive Vice President, Disney Online

"I♥ Head First HTML with CSS & XHTML—it teaches you everything you need to learn in a 'fun coated' format."

- Sally Applin, UI Designer and Artist

"Usually when reading through a book or article on design patterns, I'd have to occasionally stick myself in the eye with something just to make sure I was paying attention. Not with this book. Odd as it may sound, this book makes learning about design patterns fun.

"While other books on design patterns are saying 'Buehler... Buehler... Buehler...' this book is on the float belting out 'Shake it up, baby!""

— Eric Wuehler

"I literally love this book. In fact, I kissed this book in front of my wife."

— Satish Kumar

Other related books from O'Reilly

Statistics Hacks[™] Statistics in a Nutshell Mind Hacks[™] Mind Performance Hacks[™] Your Brain: The Missing Manual

Other books in O'Reilly's Head First series

Head First Java[™] Head First Object-Oriented Analysis and Design (OOA&D) Head First HTML with CSS and XHTML Head First Design Patterns Head First Servlets and JSP Head First EJB Head First PMP Head First SQL Head First Software Development Head First JavaScript Head First Ajax Head First Physics Head First PHP & MySQL (2008) Head First Rails (2008) Head First Web Design (2008) Head First Algebra (2008) Head First Programming (2009)

Head First Statistics



Dawn Griffiths



Beijing • Cambridge • Köln • Sebastopol • Taipei • Tokyo

Head First Statistics

by Dawn Griffiths

Copyright © 2009 O'Reilly Media, Inc. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly Media books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*safari.oreilly.com*). For more information, contact our corporate/institutional sales department: (800) 998-9938 or *corporate@oreilly.com*.

Series Creators:	Kathy Sierra, Bert Bates
Series Editor:	Brett D. McLaughlin
Editor:	Sanders Kleinfeld
Design Editor:	Louise Barr
Cover Designers:	Louise Barr, Steve Fehler
Production Editor:	Brittany Smith
Indexer:	Julie Hawks
Page Viewers:	David Griffiths, Mum and Dad

Printing History:

August 2008: First Edition.



Mum and Dad

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. The *Head First* series designations, *Head First Statistics*, and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and the authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

No snorers were harmed in the making of this book, although a horse lost its toupee at one point and suffered a minor indignity in front of the other horses. Also a snowboarder picked up a few bruises along the way, but nothing serious.

RepKover.

This book uses RepKover[™], a durable and flexible lay-flat binding.

ISBN: 978-0-596-52758-7

To David, Mum, Dad, and Carl. Thanks for the support and believing I could do it. But you'll have to wait a while for the car.

Author of Head First Statistics



Dawn Griffiths

Dawn Griffiths started life as a mathematician at a top UK university. She was awarded a First-Class Honours degree in Mathematics, but she turned down a PhD scholarship studying particularly rare breeds of differential equations when she realized people would stop talking to her at parties. Instead she pursued a career in software development, and she currently combines IT consultancy with writing and mathematics.

When Dawn's not working on Head First books, you'll find her honing her Tai Chi skills, making bobbin lace or cooking nice meals. She hasn't yet mastered the art of doing all three at the same time. She also enjoys traveling, and spending time with her lovely husband, David.

Dawn has a theory that **Head First Bobbin Lacemaking** might prove to a be a big cult hit, but she suspects that Brett and Laurie might disagree.

Table of Contents (Summary)

	Intro	xxvii
1	Visualizing Information: First Impressions	1
2	Measuring Central Tendency: The Middle Way	45
3	Measuring Spread: Power Ranges	83
4	Calculating Probabilities: Taking Chances	127
5	Discrete Probability Distributions: Manage Your Expectations	197
6	Permutations and Combinations: Making Arrangements	241
7	Geometric, Binomial, and Poisson Distributions: Keeping Things Discrete	269
8	Normal Distribution: Being Normal	325
9	Normal Distribution Part II: Beyond Normal	361
10	Using Statistical Sampling: Taking Samples	415
11	Estimating Your Population: Making Predictions	441
12	Constructing Confidence Intervals: Guessing with Confidence	487
13	Using Hypothesis Tests: Look at the Evidence	521
14	The Chi Square Distribution: There's Something Going on	567
15	Correlation and Regression: What's My Line?	605
i	Appendix i: Top Ten Things We Didn't Cover	643
ii	Appendix ii: Statistics Tables	657

Table of Contents (the real thing) Intro

Your brain on statistics. Here *you* are trying to *learn* something, while here your *brain* is doing you a favor by making sure the learning doesn't *stick*. Your brain's thinking, "Better leave room for more important things, like which wild animals to avoid and whether naked snowboarding is a bad idea." So how *do* you trick your brain into thinking that your life depends on knowing statistics?

Who is this book for?	xxviii
We know what you're thinking	xxix
Metacognition	xxxi
Bend your brain into submission	xxxiii
Read me	xxxiv
The technical review team	xxxvi
Acknowledgments	xxxvii

visualizing information **First Impressions**

Can't tell your facts from your figures?

Statistics help you make sense of confusing sets of data. They **make the complex simple**. And when you've found out what's really going on, you need a way of **visualizing** it and **telling everyone else**. So if you want to pick the best chart for the job, grab your coat, pack your best slide rule, and join us on a ride to Statsville.



Statistics are everywhere	2
But why learn statistics?	3
A tale of two charts	4
The humble pie chart	8
Bar charts can allow for more accuracy	10
Vertical bar charts	10
Horizontal bar charts	11
It's a matter of scale	12
Using frequency scales	13
Dealing with multiple sets of data	14
Categories vs. numbers	18
Dealing with grouped data	19
Make a histogram	20
Step 1: Find the bar widths	26
Step 2: Find the bar heights	27
Step 3: Draw your chart	28
Introducing cumulative frequency	34
Drawing the cumulative frequency graph	35
Choosing the right chart	39

Company Profit per Month



measuring central tendency

The Middle Way

Sometimes you just need to get to the heart of the matter.

It can be difficult to see patterns and trends in a big pile of figures, and finding the average is often the first step towards seeing the bigger picture. With averages at your disposal, you'll be able to quickly find the most representative values in your data and draw important conclusions. In this chapter, we'll look at several ways to calculate one of the most important statistics in town-mean, median, and modeand you'll start to see how to effectively summarize data as concisely and usefully as possible.









Age 19



measuring variability and spread

Power Ranges

Not everything's reliable, but how can you tell?

Averages do a great job of giving you a typical value in your data set, but they **don't tell you the full story**. OK, so you know where the center of your data is, but often the mean, median, and mode alone aren't enough information to go on when you're summarizing a data set. In this chapter, we'll show you how to take your data skills to the next level as we begin to analyze **ranges and variation**.



Wanted: one player	84
We need to compare player scores	85
Use the range to differentiate between data sets	86
The problem with outliers	89
We need to get away from outliers	91
Quartiles come to the rescue	92
The interquartile range excludes outliers	93
Quartile anatomy	94
We're not just limited to quartiles	98
So what are percentiles?	99
Box and whisker plots let you visualize ranges	100
Variability is more than just spread	104
Calculating average distances	105
We can calculate variation with the variance	106
but standard deviation is a more intuitive measure	107
Standard Deviation Exposed	108
A quicker calculation for variance	113
What if we need a baseline for comparison?	118
Use standard scores to compare values across data sets	119
Interpreting standard scores	120
Statsville All Stars win the league!	125

calculating probabilities

Taking Chances

Life is full of uncertainty.

Sometimes it can be impossible to say what will happen from one minute to the next. But certain events are more likely to occur than others, and that's where **probability theory** comes into play. Probability lets you **predict the future** by assessing how likely outcomes are, and knowing what could happen helps you make **informed decisions**. In this chapter, you'll find out more about probability and learn how to take control of the future!







Fat Dan's Grand Slam	128
Roll up for roulette!	129
What are the chances?	132
Find roulette probabilities	135
You can visualize probabilities with a Venn diagram	136
You can also add probabilities	142
Exclusive events and intersecting events	147
Problems at the intersection	148
Some more notation	149
Another unlucky spin	155
Conditions apply	156
Find conditional probabilities	157
Trees also help you calculate conditional probabilities	159
Handy hints for working with trees	161
Step 1: Finding P(Black \cap Even)	167
Step 2: Finding P(Even)	169
Step 3: Finding P(Black l Even)	170
Use the Law of Total Probability to find P(B)	172
Introducing Bayes' Theorem	173
If events affect each other, they are dependent	181
If events do not affect each other, they are independent	182
More on calculating probability for independent events	183



using discrete probability distributions Manage Your Expectations

Unlikely events happen, but what are the consequences?

So far we've looked at how probabilities tell you how likely certain events are. What probability *doesn't* tell you is the **overall impact** of these events, and what it means to you. Sure, you'll sometimes make it big on the roulette table, but is it really worth it with all the money you lose in the meantime? In this chapter, we'll show you how you can use probability to **predict long-term outcomes**, and also **measure the certainty** of these predictions.

Back at Fat Dan's Casino	198
We can compose a probability distribution for the slot machine	201
Expectation gives you a prediction of the results	204
and variance tells you about the spread of the results	205
Variances and probability distributions	206
Let's calculate the slot machine's variance	207
Fat Dan changed his prices	212
There's a linear relationship between $E(X)$ and $E(Y)$	217
Slot machine transformations	218
General formulas for linear transforms	219
Every pull of the lever is an independent observation	222
Observation shortcuts	223
New slot machine on the block	229
Add $E(X)$ and $E(Y)$ to get $E(X + Y)$	230
and subtract $E(X)$ and $E(Y)$ to get $E(X - Y)$	231
You can also add and subtract linear transformations	232
Jackpot!	238



permutations and combinations

Making Arrangements

Sometimes, order is important.

Counting **all the possible ways** in which you can order things is time consuming, but the trouble is, this sort of information is **crucial** for calculating some probabilities. In this chapter, we'll show you a **quick way** of deriving this sort of information without you having to figure out what all of the possible outcomes are. Come with us and we'll show you how to **count the possibilities**.



The Statsville Derby	242
It's a three-horse race	243
How many ways can they cross the finish line?	245
Calculate the number of arrangements	246
Going round in circles	247
It's time for the novelty race	251
Arranging by individuals is different than arranging by type	252
We need to arrange animals by type	253
Generalize a formula for arranging duplicates	254
It's time for the twenty-horse race	257
How many ways can we fill the top three positions?	258
Examining permutations	259
What if horse order doesn't matter	260
Examining combinations	261
Combination Exposed	262
Does order really matter?	262
It's the end of the race	268



geometric, binomial, and poisson distributions **Keeping Things Discrete**

Calculating probability distributions takes time.

So far we've looked at how to calculate and use probability distributions, but wouldn't it be nice to have something **easier to work with**, or just **quicker to calculate**? In this chapter, we'll show you some **special probability distributions** that follow very definite patterns. Once you know these patterns, you'll be able to use them to **calculate probabilities**, **expectations, and variances in record time**. Read on, and we'll introduce you to the geometric, binomial and Poisson distributions.

We need to find Chad's probability distribution

273



using the normal distribution

Being Normal

Discrete probability distributions can't handle every situation.

So far we've looked at probability distributions where we've been able to specify exact values, but this isn't the case for every set of data. Some types of data just **don't fit** the probability distributions we've encountered so far. In this chapter, we'll take a look at how **continuous probability distributions** work, and introduce you to one of the most important probability distributions in town—the **normal distribution**.

	(\mathbf{i})		
	000		
0		0	\bigcirc

Discrete data takes exact values	326
but not all numeric data is discrete	327
What's the delay?	328
We need a probability distribution for continuous data	329
Probability density functions can be used for continuous data	330
Probability = area	331
To calculate probability, start by finding $f(x)$	332
then find probability by finding the area	333
We've found the probability	337
Searching for a soul mate	338
Male modelling	339
The normal distribution is an "ideal" model for continuous data	340
So how do we find normal probabilities?	341
Three steps to calculating normal probabilities	342
Step 1: Determine your distribution	343
Step 2: Standardize to $N(0, 1)$	344
To standardize, first move the mean	345
then squash the width	345
Now find Z for the specific value you want to find probability for	346
Step 3: Look up the probability in your handy table	349



using the normal distribution ii

Beyond Normal

If only all probability distributions were normal.

Life can be so much *simpler* with the normal distribution. Why spend all your time working out individual probabilities when you can look up entire ranges in one swoop, and still leave time for game play? In this chapter, you'll see how to **solve more complex problems** in the blink of an eye, and you'll also find out how to bring some of that normal goodness to **other probability distributions**.

	All aboard the Love Train	363
	Normal bride + normal groom	364
	It's still just weight	365
	How's the combined weight distributed?	367
	Finding probabilities	370
	More people want the Love Train	375
	Linear transforms describe underlying changes in values	376
	and independent observations describe how many values you have	377
	Expectation and variance for independent observations	378
0	Should we play, or walk away?	383
	Normal distribution to the rescue	386
	When to approximate the binomial distribution with the normal	389
	Revisiting the normal approximation	394
	The binomial is discrete, but the normal is continuous	395
	Apply a continuity correction before calculating the approximation	396
	The Normal Distribution Exposed	404
	All aboard the Love Train	405
	When to approximate the binomial distribution with the normal	407
	A runaway success!	413



using statistical sampling

Taking Samples

10

Mighty Gumball, Inc.

Statistics deal with data, but where does it come from?

Some of the time, data's easy to collect, such as the ages of people attending a health club or the sales figures for a games company. But what about the times when data isn't so easy to collect? Sometimes the number of things we want to collect data about are so huge that it's difficult to know where to start. In this chapter, we'll take a look at how you can **effectively gather data** in the real world, in a way that's efficient, accurate, and can also save you time and money to boot. Welcome to the world of sampling.

The Mighty Gumball taste test	416
They're running out of gumballs	417
Test a gumball sample, not the whole gumball population	418
How sampling works	419
When sampling goes wrong	420
How to design a sample	422
Define your sampling frame	423
Sometimes samples can be biased	424
Sources of bias	425
How to choose your sample	430
Simple random sampling	430
How to choose a simple random sample	431
There are other types of sampling	432
We can use stratified sampling	432
or we can use cluster sampling	433
or even systematic sampling	433
Mighty Gumball has a sample	439

xix

estimating your population

Making Predictions

Wouldn't it be great if you could tell what a population was like, just by taking one sample?

Before you can claim **full sample mastery**, you need to know how to use your samples to best effect once you've collected them. This means using them to **accurately predict** what the population will be like and coming up with a way of saying how **reliable** your predictions are. In this chapter, we'll show you how knowing your sample helps you **get to know your population**, and vice versa.

\sim	•	
This is	aweson	ne!)
We hav	ve a lot (of
impres	sive sta	tistics
we can	use in o	ur
∽ adve	rtising.	r

10

40

people

prefer

pink!





constructing confidence intervals

Guessing with Confidence

12

Sometimes samples don't give quite the right result.

You've seen how you can use point estimators to estimate the **precise value** of the population mean, variance, or proportion, but the trouble is, how can you be certain that your estimate is completely accurate? After all, your assumptions about the population rely on just one sample, and what if your sample's off? In this chapter, you'll see **another way of estimating population statistics**, one that **allows for uncertainty**. Pick up your probability tables, and we'll show you the ins and outs of **confidence intervals**.

Mighty Gumball is in trouble	488
The problem with precision	489
Introducing confidence intervals	490
Four steps for finding confidence intervals	491
Step 1: Choose your population statistic	492
Step 2: Find its sampling distribution	492
Step 3: Decide on the level of confidence	494
Step 4: Find the confidence limits	496
Start by finding Z	497
Rewrite the inequality in terms of m	498
Finally, find the value of X	501
You've found the confidence interval	502
Let's summarize the steps	503
Handy shortcuts for confidence intervals	504
Step 1: Choose your population statistic	508
Step 2: Find its sampling distribution	509
Step 3: Decide on the level of confidence	512
Step 4: Find the confidence limits	513
The t-distribution vs. the normal distribution	515



using hypothesis tests Look at the Evidence

Not everything you're told is absolutely certain.

The trouble is, how do you know when what you're being told isn't right? *Hypothesis tests* give you a way of using samples to test whether or not statistical claims are likely to be true. They give you a way of *weighing the evidence* and testing whether extreme results can be explained by *mere coincidence*, or whether there are darker forces at work. Come with us on a ride through this chapter, and we'll show you how you can use hypothesis tests to confirm or allay your deepest suspicions.

Statsville's new miracle drug	522
Resolving the conflict from 50,000 feet	
The six steps for hypothesis testing	527
Step 1: Decide on the hypothesis	528
Step 2: Choose your test statistic	531
Step 3: Determine the critical region	532
Step 4: Find the p-value	535
Step 5: Is the sample result in the critical region?	537
Step 6: Make your decision	537
What if the sample size is larger?	540
Let's conduct another hypothesis test	543
Step 1: Decide on the hypotheses	543
Step 2: Choose the test statistic	544
Use the normal to approximate the binomial in our test statistic	547
Step 3: Find the critical region	548
Let's start with Type I errors	556
What about Type II errors?	557
Finding errors for SnoreCull	558
We need to find the range of values	559
Find P(Type II error)	560
Introducing power	561



the χ^2 distribution

There's Something Going On...

Sometimes things don't turn out quite the way you expect.

When you model a situation using a particular probability distribution, you have a good idea of how things are likely to turn out long-term. But what happens if there are differences between **what you expect and what you get?** How can you tell whether your discrepancies come down to normal fluctuations, or whether they're a sign of an underlying problem with your probability model instead? In this chapter, we'll show you how you can use the χ^2 distribution to **analyze your results** and sniff out **suspicious results**.

There may be trouble ahead at Fat Dan's Casino	568
Let's start with the slot machines	569
The χ^2 test assesses difference	571
So what does the test statistic represent?	572
Two main uses of the χ^2 distribution	573
v represents degrees of freedom	574
What's the significance?	575
Hypothesis testing with χ^2	576
You've solved the slot machine mystery	579
Fat Dan has another problem	585
The χ^2 distribution can test for independence	586
You can find the expected frequencies using probability	587
So what are the frequencies?	588
We still need to calculate degrees of freedom	591
Generalizing the degrees of freedom	596
And the formula is	597
You've saved the casino	599



correlation and regression

What's My Line?

Have you ever wondered how two things are connected?

So far we've looked at statistics that tell you about just one variable—like men's height, points scored by basketball players, or how long gumball flavor lasts—but there are other statistics that tell you about the **connection between variables**. Seeing how things are connected can give you a lot of information about the real world, information that you can use to your advantage. Stay with us while we show you the **key to spotting connections**: correlation and regression.



leftovers

The Top Ten Things (we didn't cover)

Even after all that, there's a bit more. There are just a few more things we think you need to know. We wouldn't feel right about ignoring them, even though they only need a brief mention. So before you put the book down, take a read through these **short but important statistics tidbits**.



#1. Other ways of presenting data	644
#2. Distribution anatomy	645
#3. Experiments	646
#4. Least square regression alternate notation	648
#5. The coefficient of determination	649
#6. Non-linear relationships	650
#7. The confidence interval for the slope of a regression line	651
#8. Sampling distributions - the difference between two means	652
#9. Sampling distributions - the difference between two proportions	653
#10. E(X) and Var(X) for continuous probability distributions	654

statistics tables

ii

Looking Things up

Where would you be without your trusty probability tables? Understanding your probability distributions isn't quite enough. For some of them, you

need to be able to **look up your probabilities** in standard **probability tables**. In this appendix you'll find tables for the **normal, t and X² distributions** so you can look up probabilities to your heart's content.



Standard normal probabilities	658
t-distribution critical values	660
γ^2 critical values	661