Life of Fred

Statistics

Stanley F. Schmidt, Ph.D.

Polka Dot Publishing

Statistics

Decisions! Decisions! Decisions! Do you attend Harvard University or KITTENS University? Do you marry this person or not? Does your pizza company continue the television advertising which features the "Pizza for People Who Like to Canoodle" slogan?

Success in life is 90% making the right decisions in the first place. And only 10% carrying out those decisions.

People with good decision-making skills are rare. They are also the most valuable persons in any business, any army or any orchestra. These CEOs, generals and conductors all have the same job: they take massive amounts of data and boil them down to yes-or-no decisions.

✓ Shall we sell all the stock we own? (It's September 1929.)
✓ Shall we launch the invasion today? (It's June 6, 1944.)
✓ Shall we send our orchestra on a worldwide tour this month? (It's early December 1941.)

And where there are numbers involved, statistics is an important aid in making good decisions. At its best, statistics is a way of melting down a heap of numerical data into a simple yes or no. *It's a way of getting rid of numbers!*

If you really hate to see big piles of numbers, you and statistics were made for each other.

A Note to Students

One morning in the life of Fred. A Saturday just after his sixth birthday. In his everyday life Fred will run into the need for every kind of statistics. Each time we do a little statistics, we see how it helps him get through his morning.

HOW MUCH STATISTICS IS COVERED IN THIS BOOK?

We start at the beginning with simple descriptive statistics (averages, standard deviation, etc.) and then do some probability, including conditional probability with Bayes' Theorem.

Next comes inferential statistics—the heart of statistics—in which we study a zillion* different procedures. We describe each in detail and tell you when and where each test is appropriate. You get plenty of worked-out examples for each test.

All the popular tests such as the Normal Distribution and the Chi-Squared Test are included. Many advanced tests such as the Kolmogorov-Smirnov Test and the Two-Factor ANOVA for multiple observations per cell are covered. When the Chi-Squared Test won't work because the sample sizes are too small, we turn to Fisher's Exact Test. Most beginning statistics books don't include that test.

We have one test *that no other statistics book mentions*—at least not until future authors copy it out of this book. It deals with A SMALL SAMPLE FROM A BINOMIAL DISTRIBUTION. Suppose, for example, a new species of fish is discovered in the ocean and of the first ten caught, three had red fins. What is the number of red-finned fish you might expect if you caught 10,000? [Answer: 95% of the time, you would expect between 1093 and 6096.] This question would stump most statistics teachers (who don't have a copy of this book).

After the descriptive and inferential statistics, we spend the last hour or so of Fred's morning working with regression equations including nonlinear curve fitting and logistic regression.

This book has much more material than is normally covered in a beginning university statistics course.

✶ 46 by actual count

HOW THEORETICAL IS THIS BOOK?

Life is practical. This is a book that will teach you how to *do* statistics—lots of it. Even if you are going to get a Ph.D. in statistics and are dying to go through tons of theory and proofs, your first logical step should be to learn how to do the various tests. Then, in a later course, the proofs would be appropriate. In beginning algebra, for example, you were first told that a negative number times a negative number gives a positive answer. Later, you may have seen the proof.

In this book you learn how to perform the Kruskal-Wallis Test for three or more independent samples, but we're not going to fill up the pages with a proof.

There are two exceptions. The first is a little three-line proof of Bayes' Theorem, which is so cute that I couldn't resist including it. And the second is the underpinnings of the SMALL SAMPLE FROM A BINOMIAL DISTRIBUTION TEST that I mentioned on the previous page. Since no other book has this test, I placed this material in its own separate little chapter (Chapter 5½) and laid out the reasoning to show why this test works. This little chapter is the only place in the book in which there is any calculus. And even here, the calculus is very basic. It deals with the area under a curve described by a polynomial. If you go directly from Chapter 5 to Chapter 6 and bypass Chapter 5½, you will be protected from all calculus.

In doing their proofs, some books go nuts with subscripts and primes and "hats" and Greek letters. They wind up with expressions like $\hat{y}'_{i,j} + \epsilon$, which certainly don't help anyone's digestion. Those things are kept to a minimum in *Life of Fred: Statistics*. [$\hat{y}$ is read "y-hat."]

WHAT BACKGROUND DO I NEED?

It would be nice to have a little algebra so that x^2 and absolute values and square roots don't mystify you. But that's about it. I can't think of anywhere in the book where you'll need to solve any equations or do any algebra word problems.*

We'll use the greater than sign (>) and plus-or-minus (±).

---

* None of those old word problems like: JACKIE IS CHASING DALE DOWN THE HALL WITH AN AX. JACKIE IS TRAVELING 7 FT/SEC AND DALE IS RUNNING AT 5 FT/SEC. THEY ARE 8 FEET APART. HOW SOON SHOULD DALE START APOLOGIZING?

See if these all make sense to you:

- ☞ $7^2 = 49$
- ☞ $|-3| = 3$
- ☞ $64 > 29$
- ☞ 7 ± 2 means 5 or 9.
- ☞ Using your calculator $\sqrt{3}$ gives `1.7320508`.

If so, you are ready.

DO I NEED A COMPUTER?

No.

DO I NEED A GRAPHING CALCULATOR?

No. All you need is a handheld calculator that has keys like sin, cos, and log. Those calculators don't cost that much. Certainly under $20. (I have seen them under $10.) In a couple of years they will probably be included free in cereal boxes.

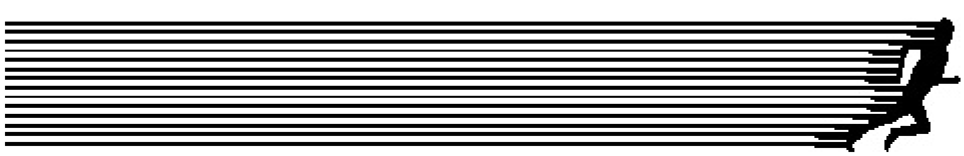
ANY SPECIAL SUGGESTIONS BEFORE I START CHAPTER ONE?

Yes. I have a couple of ideas.

First, in each chapter there are *Your Turn to Play* sections. These have representative problems along with completely worked-out solutions. Please solve these problems before you glance at the solutions. Just reading the problems and eyeballing the solutions is a real temptation for some readers, but unless you're smarter than Einstein, you won't learn much doing that.

At the end of each chapter are six sets of exercises which I call Cities. The first two Cities have all the answers supplied. The second pair of Cities have answers to all the odd-numbered problems. If you want to learn statistics, please do all these problems for which I have given the answers.

Second, I need to know if you are in a real hurry.



If that's the case, then don't start by turning to the first page of Chapter One, or to the Table of Contents or to the Index.

Instead, turn to the Emergency Statistics Guide which begins on p. 326. The Emergency Statistics Guide will tell you:

① what test to use,

② where to find an explanation of the test as it occurred in Fred's life,

③ where it's listed in the Field Guide and

④ what table to use.

The Emergency Statistics Guide will move you from baffled to brilliant in twelve seconds flat.

Contents

Exponential Distribution
Normal Distribution

Past—The sample is known and you want to know what the population *was* that gave this sample. You start with one sample.

Normal Distribution—$n > 30$ and the sample is small compared with the population.
Normal Distribution—$n > 30$ and the sample is large compared with the population.
Student's *t*-Distribution
Binomial Distribution (large sample, $n > 30$)
Binomial Distribution (small sample, $n \leq 30$)
Kolmogorov-Smirnov goodness-of-fit test
Lilliefors test
Wilcoxon Signed Ranks test
Sign test—Does the population have that median?
Sign test for Nominal Data
Chi-Squared test (goodness of fit)
Chi-Squared test (Lie Detector)
Chi-Squared test (Is the population too variable?)
Runs test

Present—You start with two samples and want to know how *do* they compare with each other.

Two Paired Samples ($\mu_1 - \mu_2$)
Wilcoxon Signed Ranks test
Sign test for two paired samples
Sign test for two paired samples of nominal data.
Two Proportions in two categories.
Two Large Independent Samples, $n \geq 30$
Two Independent Samples (σ_1 and σ_2 known)
F-Distribution test
Two Small Independent Samples, roughly equal standard deviations
Two Small Independent Samples (Smith-Satterthwaite) with very different standard deviations.
Mann-Whitney test (a.k.a. Wilcoxon Rank-Sum test)
Chi-Squared test (χ^2), two samples of nominal data in multiple categories.
One Sample with Two Variables
Chi-Squared test (χ^2)—Yates Correction
Fisher's Exact test

Present—You start with three or more samples and want to know how *do* they compare with each other.

One-Way ANOVA (independent samples)
Post-test for One-Way ANOVA (independent samples)

Chapter One
Descriptive Statistics

Tink! Fred's eyes popped open. He had just heard one of the sweetest sounds. He looked at his watch. 4:13 A.M. With his mouth open, he listened in the dark. Tink! Yes, he thought to himself, it's happened. Tink!

Drops of water were falling from the ceiling. Fred threw off his bedcovers and emerged from under his desk. He looked at the pot on his desktop and saw three drops of water. Tink! Make that four drops.

His watch clicked over to 4:14 A.M. and he smiled as six more drops fell into the pot. It's a little early to telephone Alexander, Fred thought, but it won't hurt if I email him. Fred rolled up his three-foot sleeping bag and put it in the closet. He turned on the computer, changed out of his pajamas, turned off his nightlight, and looked out the window. From the window in his office/home he could look out over the university campus. For the first time in months, the sky was inky black and filled with stars. It was a welcome change from what he called the "dodo bird" sky of Kansas in winter.

From September through May, the cloud cover always reminded Fred of the soft, gray feathers of that extinct bird.

He opened the window and felt a warm breeze. So much to be grateful for. I teach at a wonderful university. I have my health. I have wonderful friends like Alexander and Betty. Fred uttered the prayer that God most likes to hear ("Thank you") and then turned to his computer that was in the final stages of booting up. He put three phone books on a chair and hopped on top of them. When you're only six years old and 36 inches tall, you need to make those kinds of adjustments in order to sit at a big-people's desk.

On a clipboard he wrote out a little **frequency distribution** showing the data he had collected so far:

time	no. of drops
4:13	4
4:14	6

But that looked much too "numbery" for Fred's taste. He liked to keep things simple. Instead of

4:13 A.M., Fred wrote "1" to stand for the first minute of spring, and "2" for the second minute.

His frequency distribution looked much nicer now:

time	no. of drops
1	4
2	6

He stared at the computer screen. Three operating systems had been loaded, the anti-virus program and the anti-spam programs were activated, and the screen colors were being adjusted to match the university colors, and now the Internet service provider was being dialed.

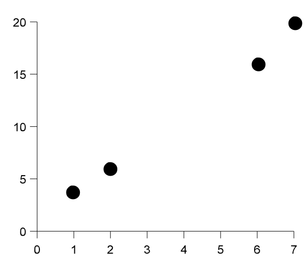
Fred had a very new machine (it was a gift from his students), but the university had very old phone lines. "ISP IS NOT RESPONDING" appeared on his screen. "ERROR 397 THE NUMBER IS BEING REDIALED."

Fred went back to looking at the pot. It was 4:19 A.M. and during that minute Fred counted 16 drops coming from the ceiling into his pot. His screen flashed, "LOCAL NUMBER IS UNAVAILABLE. THE NEVADA NUMBER IS BEING DIALED." Fred went back to counting. Twenty drops came in the next minute. "THE NEVADA NUMBER IS BUSY. URUGUAY IS BEING DIALED."

Fred went back to his clipboard and expanded his frequency distribution:

time	no. of drops
1	4
2	6
6	16
7	20

To pass the time waiting for his computer, he drew a little graph.



Fred's Scatter Diagram

A bunch of dots on a graph (where paired observations are plotted) is called a **scatter diagram**.)

Quick! No time to wade through a table of contents or an index. Do you use the Kolmogorov-Smirnov one-sample test or do you perform a Chi-Squared test? Should you resort to the Wilcoxon Signed Ranks test? Or two-factor ANOVA?

Just answer these questions and follow the arrows. You'll learn exactly which statistics procedure you'll need.

Do you really want to mess with the numbers and formulas yourself?

NO → *Hiring a statistician is an alternative. Please have your checkbook ready.*

YES ↓

You have a population and you have sample(s) drawn from that population.

✳ You might be just starting with a population and are wondering what samples from that population will look like. → *Go to next page.*

✳ You have a single sample and you want to describe the population from which it might have come. → *Turn to page 329.*

✳ You have two samples and you want to compare them. → *Turn to page 333.*

✳ You have three or more samples and you want to see how they fit together. → *Turn to page 339.*

From the previous page.

You have a population and you want to know what samples from this population *will* look like.

We call this the Future.

You are starting with zero samples.

Are the data nominal/ordinal or are they interval/ratio?
(The types of data—nominal, ordinal, interval and ratio—are all explained on p. 176)

interval or ratio →

Normal Distribution

Explanation . . p. 116
Field Guide . . . p. 401
Table . . . . p. 470

↓ nominal or ordinal data

Are you dealing with "hits" scattered randomly throughout an interval (of time, distance, etc.)
or
is it more like drawing colored marbles out of a jar?

marbles in a jar →

Go to next page.

↓ hits in an interval

Do you want to find the probability of getting g hits in the next five minutes (or five miles or five gallons or five acres) →

Poisson Distribution

Explanation . . . . p. 95
Field Guide . . . . p. 399

or

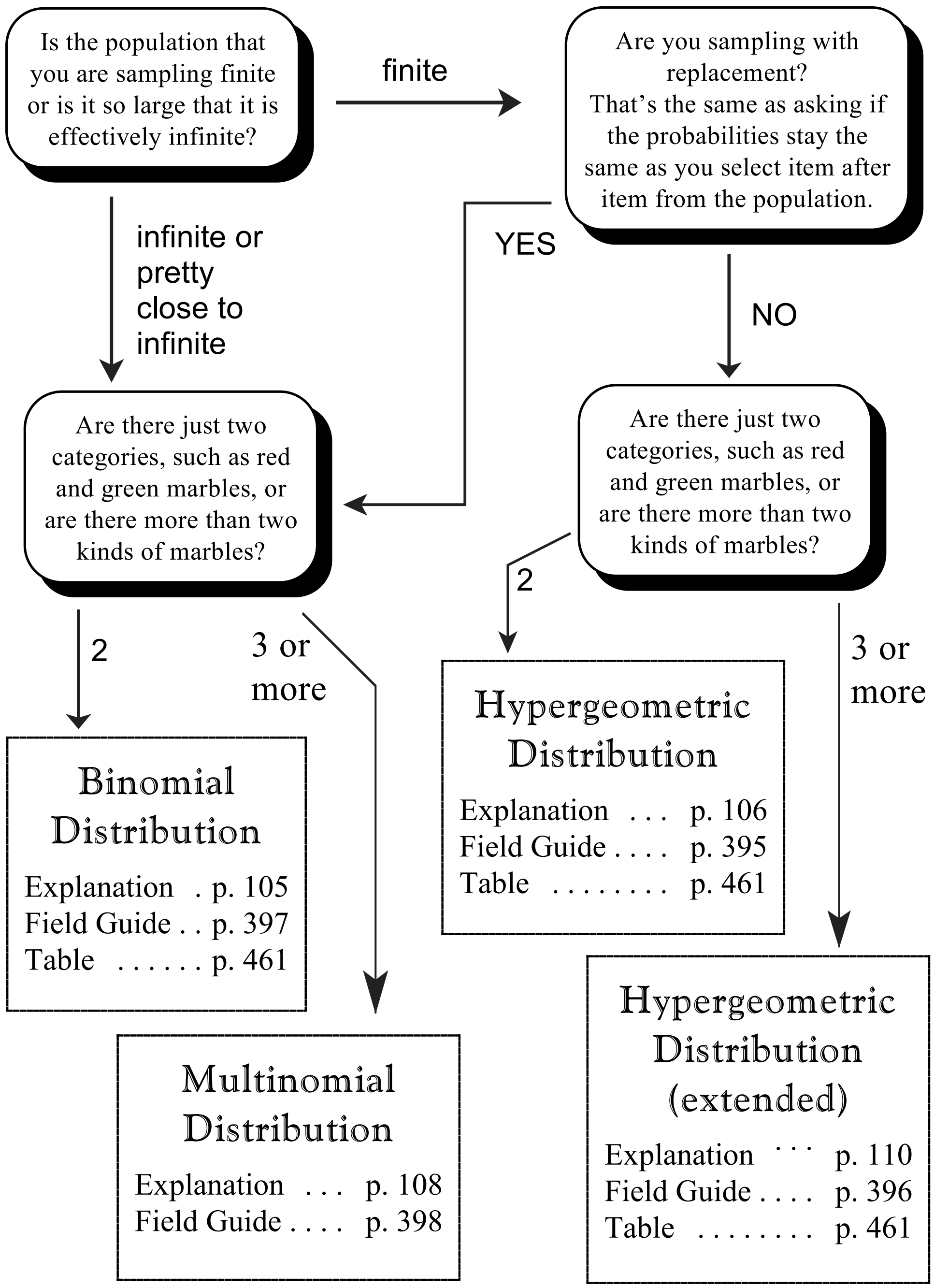
do you want to find the probability of getting the next hit in the next five minutes? →

Exponential Distribution

Explanation . . . . p. 100
Field Guide . . . . p. 400

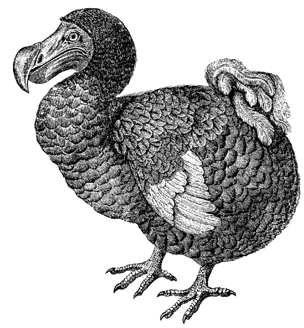
From the previous page.

You are drawing colored marbles out of a jar.



Field Guide

Field guides can be fun. When you're out tromping in the woods and spot this large, flightless bird. You want to know what it is. Your hiking companion remarks, "It ain't a chicken, but it looks like good eating."



dodo bird

You pull out your *Acme Field Guide to Birdies* and read aloud, "RAPHUS CUCULLATUS: length is about three feet. Soft, gray feathers. Forest-dwelling. Female lays one egg in a bed of grass. Used to live in Mauritius."

"That's fine," your friend says as he takes aim with a rock. "That thing can't even fly. Did you bring your cookbook?"

"Stop!" you exclaim. "We really don't want to kill that dodo. My field guide says that they became extinct in 1693. We could make a zillion dollars if we capture this guy alive."

See how handy a field guide can be?

Mauitius

And if you had a field guide for geography, you could locate Mauritius. It's east of Madagascar. And Madagascar is a large island in the Indian Ocean. And the Indian Ocean is east of the southern part of Africa.

All of which forms the perfect segue (SEG-way, a smooth transition) to our Field Guide to Statistics. Our "birds" are divided into three different types. These three types of statistical maneuvers are called the Future, the Past, and the Present. Each of them will answer a particular type of question.

The Future begins with a known population. With these procedures we will be able to predict what a sample from this population *will* look like.

The Past begins with a given sample. We look at the sample and try to describe what *was* the population that generated that sample.

The Present begins with two or more samples and learns how these samples *fit* in the general scheme of things.

Fisher's Exact Test

Type of Test
- ☐ "Future"—Population known. What *will* the sample be like?
- ☐ "Past"—Sample known. What *was* the population that gave this sample?
- ☒ "Present"—Multiple samples. How *do* they fit?

Type of data
- ☒ Nominal—3 wins, 2 loses
- ☐ Ordinal—freshman, sophomore, junior, senior
- ☐ Interval—temperature
- ☐ Ratio—$6, $7, $23

Special Features

Fisher is used when the Chi-Squared won't work because the samples are too small—even after you have combined samples and categories down to a 2 × 2 contingency table.

Variables You Know

In attempting to use the χ^2 test, you have combined samples and categories and still have too few items of data in a 2 × 2 contingency table:

w	x
y	z

w, x, y and z are numbers

What You Can Find Out

What is the probability that the two populations are alike?

The Procedure

Step1: Interchange the rows and/or interchange the columns until the number in the upper left-hand box is the smallest of the four numbers.

Step 2: If the number in the upper right-hand box is greater than the number in the lower left-hand box, interchange the two numbers.

Step 3: Your contingency table will now look like

A	B
C	D

where $A \le B \le C$ and $A \le D$.

With these values of A, B, C and D, consult Table M (on p. 499) for the exact probability that the populations are the same.

Example *In competition, Stanthony's Veal Pizza has won 6 Grand awards and lost 3 times. His Linguini-and-Antelope Pizza has won 2 Grand awards and lost 8 times. At the 5% significance level are these two pizzas different in their likelihood of winning awards?*

$$\begin{matrix} 6 & 3 \\ 2 & 8 \end{matrix} \quad \text{interchanging rows} \quad \begin{matrix} 2 & 8 \\ 6 & 3 \end{matrix} \quad \text{interchanging upper right/lower left} \quad \begin{matrix} 2 & 6 \\ 8 & 3 \end{matrix}$$

So A = 2, B = 6, C = 8, and D = 3. Table M states there is a 0.0549 probability that the samples were drawn from the same population. That exceeds 5%. At the 95% confidence level we could not say that the two pizzas differed in their likelihood of winning awards.

Index