CSSS 2000–2001 Math Review Lectures: Probability, Statistics and Stochastic Processes

Cosma Rohilla Shalizi
Physics Department, University of Wisconsin-Madison, and the Santa Fe Institute[1]

Presented 4 June 2000
and 11 June 2001

[1]Present address: Statistics Department, Carnegie Mellon University, cshalizi@cmu.edu

These notes are available electronically from `http://bactra.org/prob-notes/`.
Reports of typos, and more serious bugs, are eagerly requested.

Revision History

Date	Comments
4 June 2000	First version finished (about four hours before teaching from it).
8 July 2000	Minor typo fixes, some points amplified in response to student comments; added list of deficiencies as Appendix B.
11 June 2001	More typo fixes, added paragraph on large deviation principle, second presentation.
14 July 2001	Fixed two errors (thanks to Cecile Viboud and Antonio Ramirez)
2 January 2006	Typo fix (thanks to Mario Negrello)
8 February 2006	Typo fix (thanks to Julia Uddén)
12 April 2006	Typo fix (thanks to Raouf Ghomrasni)

Contents

Part I: Probability **1**

1 What's a Probability Anyway? **1**

2 Probability Calculus **2**
2.1 Basic Rules . . . 2
2.1.1 A Caution about Probabilities 0 and 1 . . . 3
2.2 Conditional Probabilities . . . 3

3 Random Variables **5**
3.1 Properties of Random Variables . . . 6
3.1.1 Functions of Random Variables . . . 6
3.1.2 Multiple Random Variables; Independence . . . 6
3.2 Expectation . . . 6
3.2.1 Expectations of Multiple Variables . . . 7
3.3 Moments . . . 7
3.3.1 Particularly Important Moments . . . 8
Mean . . . 8
Variance and Standard Deviation . . . 8

4 Important Discrete Distributions **9**
4.1 The Bernoulli Distribution . . . 9
4.2 The Binomial Distribution . . . 9
4.3 The Poisson Distribution . . . 10
4.4 First Moments of These Distributions . . . 10

5 Continuous Random Variables **11**
5.1 The Cumulative Distribution Function . . . 11
5.2 The Probability Density Function . . . 11
5.3 Continuous Expectations . . . 12

6 Important Continuous Distributions **13**
6.1 The Exponential Distribution . . . 13
6.2 The Normal or Gaussian Distribution . . . 13
6.3 The χ^2 Distribution . . . 14

6.4 The Lognormal Distribution 14
6.5 Power-Law Distributions 14
6.6 First Moments and pdfs of These Distributions 15

7 Tricks with Random Variables 16
7.1 The Law of Large Numbers 16
7.2 The Central Limit Theorem 17
7.2.1 The Extraordinary Importance of the CLT 17
7.2.2 Ways in Which the CLT Can Fail 18
7.2.3 The Many-Independent-Causes Story for Why Things Are Gaussian 18
7.2.4 Multiplicative Noise and the Lognormal Distribution . . . 19

Part II: Statistics 20

8 The Care and Handling of Data 21
8.1 Counting 21
8.1.1 The Question of Bins 21
8.1.2 Histograms 22
8.1.3 Percentiles 22
8.1.4 Median 22
8.1.5 Mode 22
8.2 Adding 22
8.2.1 Sample Mean 22
8.2.2 Sample Variance 23
8.3 Correlating 23
8.3.1 Covariance 23
8.3.2 The Correlation Coefficient 23

9 Sampling 25
9.1 The Notion of "A Statistic" 25
9.2 Loss of Variability Under Sampling 25
9.3 Figuring the Sample Distribution; Monte Carlo Methods 26

10 Estimation 27
10.1 Point Estimates 27
10.1.1 Bias, Variance, and Other Sorts of Quality 27
10.1.2 Some Common Kinds of Estimates 28
Least Squares 28
Maximum Likelihood 29
10.2 Curve-Fitting or Regression 30
10.3 Propagation of Errors 30
10.4 Confidence Regions 31

11 Hypothesis Testing **33**
11.1 Goodness-of-Fit . . . 33
11.1.1 Significant Lack of Fit . . . 33
11.1.2 The χ^2 Test . . . 34
11.2 The Null Hypothesis and Its Rivals . . . 35
11.2.1 The *Status Quo* Null . . . 35
11.2.2 The It-Would-Be-the-Worst-Mistake Null . . . 35
11.2.3 The Random-Effects Null . . . 36
11.2.4 The Alternative Hypothesis . . . 36
11.3 Formalism of Tests . . . 37
11.3.1 The Test Statistic . . . 37
11.3.2 The Regions . . . 37
11.3.3 The Kinds of Errors; Error Probabilities . . . 37
Significance Level or Size . . . 37
Power . . . 38
Severity . . . 38
The Trade-Offs . . . 39
11.3.4 Test for Whether Two Sample Means Are Equal . . . 39

12 Funky Statistics **41**
12.1 Nonparametric Estimation and Fitting . . . 41
12.2 Machine Learning . . . 41
12.3 Causal Inference . . . 42
12.4 Ecological Inference . . . 42
12.5 Optimal Experimental Design . . . 42

Part III: Stochastic Processes **43**

13 Sequences of Random Variables **43**
13.1 Representing Stochastic Processes with Operators . . . 44
13.2 Important Properties of Stochastic Processes . . . 44
13.2.1 Stationarity . . . 44
13.2.2 Ergodicity . . . 45
13.2.3 Mixing . . . 45

14 Markov Processes **46**
14.1 Markov Chains and Matrices . . . 46
14.2 Some Classifications of States, Distributions and Chains . . . 47
14.3 Higher-Order Markov Chains . . . 48
14.4 Hidden Markov Models . . . 48

15 Examples of Markov Processes **49**
15.1 Bernoulli Trials . . . 49
15.2 Biased Drift on a Ring . . . 49
15.3 The Random Walk . . . 49

16 Continuous-Time Stochastic Processes **51**
16.1 The Poisson Process 51
16.1.1 Uses 52
16.2 Brownian Motion, or the Wiener Process 52

A Notes for Further Reading **54**
A.1 Probability 54
A.2 Statistics 54
A.3 Stochastic Processes 55

B What's Wrong with These Notes **57**

Chapter 1

What's a Probability Anyway?

As far as pure probability theory is concerned, probabilities are real numbers between 0 and 1, attached to sets in some mathematical space, assigned in a way which let us prove nifty theorems. We're not going to worry about any of these details.

Mathematical probabilities make good models of the frequencies with which events occur, somewhat in the same way that Euclidean geometry makes a pretty good model of actual space. The idea is that we have a space of occurrences which interest us — our **probability space**. We carve this up into (generally overlapping) sets, which we call **events**. Pick out your favorite event A, and keep track how often occurrences in A happen, as a proportion of the total number of occurrences; this is the **frequency** of A. In an incredibly wide range of circumstances, frequencies come very close to obeying the rules for mathematical probabilities, and they generally come closer and closer the longer we let the system run. So we say that the probability of A, P(A), is the limiting value of the frequency of A. [1]

Take-home: *probabilities are numbers which tell us how often things happen.*

[1] The foundations of probability are one of the most acrimoniously disputed topics in mathematics and natural science; what I'm spouting here is pretty orthodox among stochastic process people, and follows my own prejudices. The main alternative to the "frequentist" line is thinking that probabilities tell you how much you should believe in different notions. This "subjectivist" position is OK with assigning a probability to — say — the proposition that a perpetual motion machine will be constructed within a year.

Chapter 2

Probability Calculus

The function which takes an event and gives us its probability is called the **probability distribution**, the **probability measure**, or simply the **distribution**. It generally isn't defined for *every* possible subset of the probability space; the ones for which it is defined, the "good" (technically: measured) events, are sometimes called the "field" of the distribution. We won't need that terminology, but it's good not to be frightened of it when you run across it.

2.1 Basic Rules

Here A and B are any two events. Following custom, Ω is the special event which contains every point in our probability space, and $\emptyset$ is the event which contains no points, the empty set. $\bar{\text{A}}$ is the **complement** of A, the event which is all the occurrences which are not in A. A+B is the union of A and B; AB is the intersection of A and B. (Both A+B and AB are also events.)

1. $0 \leq \text{P}(\text{A}) \leq 1$ (Events range from never happening to always happening)
2. $\text{P}(\Omega) = 1$ (*Something* must happen)
3. $\text{P}(\emptyset) = 0$ (Nothing never happens)
4. $\text{P}(\text{A}) + \text{P}(\bar{\text{A}}) = 1$ (A must either happen or not-happen)
5. $\text{P}(\text{A} + \text{B}) = \text{P}(\text{A}) + \text{P}(\text{B}) - \text{P}(\text{AB})$

The last rule could use a little elaboration. The meaning of A+B is "A alone, or B alone, or both together". To figure out how often it happens, we add how often A and B happen ($\text{P}(\text{A}) + \text{P}(\text{B})$) — but *each* of those includes A and B happening together, so we're counting those occurrences twice, and need to subtract P(AB) to get the right value.

What follows are some simple exercises which give you useful rules for manipulating probabilities. Some of them should be trivial, but do the exercises anyway.

Exercise. Convince yourself that $\mathrm{P}(\mathrm{AB}) \leq \mathrm{P}(\mathrm{A})$.

Exercise. Convince yourself that $\mathrm{P}(\mathrm{A}+\mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$ if A and B are mutually exclusive or **disjoint**.

Exercise. Convince yourself that if $\mathrm{A}_0, \mathrm{A}_1, \ldots \mathrm{A}_n$ are mutually exclusive and jointly exhaustive, then $\mathrm{P}(\mathrm{A}_0) = 1 - \sum_{i=1}^{n} \mathrm{P}(\mathrm{A}_i)$.

Exercise. Use (4) and (5) to show (2); or use (2) and (5) to show (4).

2.1.1 A Caution about Probabilities 0 and 1

Ω is not necessarily the only event with probability 1, nor $\emptyset$ the only one with probability zero. In general, probability 0 means that an event happens so rarely that in the limit we can ignore it, but that doesn't mean it *never* happens. Probability 1 events are said to happen "almost always", or "almost surely" (abbreviated **a.s.**) or "almost everywhere" (**a.e.**), while probability 0 events happen "almost never".

Exercise. Convince yourself that if there is an event $\mathrm{A} \neq \emptyset$ for which $\mathrm{P}(\mathrm{A}) = 0$, then there are events smaller than Ω with probability one.

2.2 Conditional Probabilities

Suppose you're interested only in part of the probability space, the part where you know some event — call it A — has happened, and you want to know how likely it is that various other events — B for starters — have also happened. What you want is the **conditional probability** of B given A. We write this $\mathrm{P}(\mathrm{B}|\mathrm{A})$, pronouncing the vertical bar | as "conditioned on" or "given". We can write this in terms of unconditional probabilities:

$$\mathrm{P}(\mathrm{B}|\mathrm{A}) \equiv \frac{\mathrm{P}(\mathrm{AB})}{\mathrm{P}(\mathrm{A})} , \tag{2.1}$$

which makes some sense if you stare at it long enough.

Conditional probabilities are probabilities, and inherit all the necessary properties; just re-write (1)–(5) above with bars and extra letters in the right place.[1] (Get used to seeing the bars.)

If $\mathrm{P}(\mathrm{B}|\mathrm{A}) = \mathrm{P}(\mathrm{B})$, then whether or not A happens makes no difference to whether B happens. A and B are then said to be **independent** or **statistically independent**. (If B is independent of A, then A is independent of B. *Exercise:* show this.) It is often extremely useful to break probability problems up into statistically independent chunks, because there's a lot of machinery for proving results about those.

[1] You may be worrying about what happens when $\mathrm{P}(\mathrm{A}) = 0$. That is one of the conditions under which conditional probabilities can fail to exist — but sometimes they're mathematically well-defined even when the event we condition on has zero probability. If you *really* want to worry about this, read Billingsley (1995).

Conditional probabilities can be inverted. That is,

$$\mathrm{P(A|B)} \quad = \quad \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}} \ . \tag{2.2}$$

This relationship is called **Bayes's Rule,** after the Rev. Mr. Thomas Bayes (1702–1761), who did not discover it.

Exercise. Prove Bayes's Rule from the definition of conditional probability.

Exercise. Suppose $\mathrm{A}_0, \mathrm{A}_1, \ldots \mathrm{A}_n$ are mutually exclusive and jointly exhaustive events. Prove the following form of Bayes's Rule:

$$\mathrm{P}(\mathrm{A}_i|\mathrm{B}) \quad = \quad \frac{\mathrm{P(B|A_i)P(A_i)}}{\sum_{j=1}^{n} \mathrm{P(B|A_j)P(A_j)}} \tag{2.3}$$

Exercise. Show that A and B are independent if and only if P(AB) = P(A)P(B).

A and B are said to be **conditionally independent** given C (or **independent conditional on** C) when P(AB|C) = P(A|C)P(B|C).

Exercise. If A and B are independent, are they still necessarily independent when conditioning on any set C?

Chapter 3

Random Variables

Essentially anything which has a decent probability distribution can be a random variable. A little more formally, any function of a probability space, $f : \Omega \mapsto \Xi$, is a random variable, and turns its range (the space Ξ) into a probability space in turn. So start with your favorite abstract probability space Ω, and take functions of it, and functions of functions, until you come to what you need, and you have so many random variables.[1]

It's conventional to write random variables with upper-case italic letters: A, B, C, X, Y, Z, and so forth. (Note that I've been writing events as upper-case romans, A, B, C.) We write the particular values they may take on — **realizations** — in lower-case italics: a, b, c, x, y, z. We say "The roll of this die is a random variable, X; this time it came up five, so $x = 5$." You'll have to trust me that this leads to clarity in the long run.

If X is a random variable, then the space Ξ it lives in must be a probability space, and that means there needs to be a distribution over Ξ. That distribution is fixed by the distribution on the original space Ω and the mapping from Ω to Ξ. We simply define the probability of a set $\mathrm{A} \subset \Xi$ as $\mathrm{P}(f^{-1}(\mathrm{A}) \subset \Omega)$. That is, the event A, in the space Ξ, is equivalent to the set of points in Ω which map to A under f, and we know the probability of that.[2] We say that the distribution on Ω and the function f **induce** a distribution on Ξ.

[1] There are actually ways of defining "random" which don't invoke probability, just algorithms, so that the original abstract space can be constructed legitimately. This is a very interesting and strange topic which will probably be covered by other instructors.

As a further aside, if you know some category theory (Cohn 1981), you can set up a category of random variables, where the objects are probability spaces and the morphisms are measurable functions.

[2] As usual, this ignores measure-theoretic quibbles about whether $f^{-1}(\mathrm{A})$ is a proper event, i.e., whether that set is measurable. The more elegant way of avoiding this difficulty is to confine our attention to sets in Ξ whose pre-images in Ω are, in fact, measurable.

3.1 Properties of Random Variables

3.1.1 Functions of Random Variables

Any function of a random variable is a random variable. As with random variables themselves, the distribution of the values of that function is said to be the **induced** distribution.

Notice that the above includes constant functions of random variables. This is a bit of a degenerate case, but it'll prove to be useful mathematically.

3.1.2 Multiple Random Variables; Independence

Given *two* random variables, we can construct a larger probability space in which each variable provides one coordinate. The distribution over this larger space is called the **joint distribution**, and we usually write it like $\mathrm{P}(X = x, Y = y)$, meaning the probability of getting $X = x$ and $Y = y$ at the same time. Two random variables are **independent** when $\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y)$ for all x and y. The **conditional distribution** is defined in the natural way,

$$\mathrm{P}(X = x|Y = y) \equiv \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(Y = y)} . \tag{3.1}$$

The **marginal distribution** of one variable say X, is just $\mathrm{P}(X = x)$. You can think of it as $\mathrm{P}(X = x|Y = y)$ averaged over all values of y. (See below, on expectations.)

If X and Y are independent, then functions of them are independent of each other: i.e. for any functions f, g, $f(X)$ is independent of Y and $g(Y)$, and $g(Y)$ is independent of X.

All of this extends in the natural way to three or more random variables.

Exercise. Convince yourself that X and Y are independent when $\mathrm{P}(X = x|Y = y) = \mathrm{P}(X = x)$.

3.2 Expectation

So we have our probability space Ω, and we have a probability distribution P over it, and together they give us a random variable X. Now imagine that we have a function f which takes a point in the space and outputs a number — it could be an integer or a real or complex or some vector, so long as we can add them up and multiply them by real numbers. So now we want to know, if we pick inputs to f according to our distribution, what will the average value of the output be? This is called the **expectation value** of f, or just its **expectation** for short. It's written as $Ef(X)$, or $\mathbf{E}f(X)$, or $E\{f(X)\}$, or $Exp\{f(X)\}$, or, especially in physics, as $f(X)$ or $\langle f(X)\rangle$. (Sometimes the argument to f is dropped if it's clear what random variable we're talking about. Sometimes it's dropped even if it isn't clear, alas.) I'll use $\mathbf{E}f(X)$, and sometimes $f(\bar{X})$, simply because they're easy to write in TEX, and keep it clear that taking expectations

is an operation we perform on the function, that we're not multiplying it by some number E; for the same reason, I'll use $E\{f(X)\}$ at the blackboard.

Assuming, as we are for now, that Ω is discrete, so that every point has a probability assigned to it, then we can write

$$\mathbf{E}f(X) \equiv \sum_{x\in\Omega} f(x)\mathrm{P}(X=x) \; . \tag{3.2}$$

That is, the expectation value is just a weighted sum of the values of $f(x)$, the weights being given by $\mathrm{P}(X=x)$.

Taking expectations is a linear operator: $\mathbf{E}(f(X)+g(X)) = \mathbf{E}f(X)+\mathbf{E}g(X)$, and $\mathbf{E}\alpha f(X) = \alpha\mathbf{E}f(X)$.

Exercise. Convince yourself of this from the defining formula.

We will sometimes need to take expectations using several different distributions. To avoid confusion, expectation with respect to distribution θ will be written $\mathbf{E}_\theta$.

3.2.1 Expectations of Multiple Variables

If we have a function of two or more random variables, it's a random variable itself of course, and we can take its expectation in the obvious way. If $Z = f(X,Y)$,

$$\mathbf{E}Z \equiv \sum_{x,y} f(x,y)\mathrm{P}(X=x, Y=y) \; . \tag{3.3}$$

From the defining formula, $\mathbf{E}(f(X)+g(Y)) = \mathbf{E}f(X)+\mathbf{E}g(Y)$. If If X and Y are independent, then $\mathbf{E}f(X)g(Y) = (\mathbf{E}f(X))(\mathbf{E}g(Y))$ as well.

Exercise. Convince yourself of the statement about addition of random variables from the definition.

Exercise. Convince yourself of the statement about multiplying independent random variables.

Conditional expectations are just expectations taken with respect to some conditional probability distribution or other. To indicate the expectation of X conditioned on $Y=y$, we write $\mathbf{E}(X|Y=y)$.

Fun but irrelevant fact: It's possible to define probabilities in terms of expectations, if you start with certain axioms about them. Some people find this comforting.

3.3 Moments

The **moments** of a distribution are the expectations of various powers of its random variable. That is, we assume that the points in our probability space are things we can add up and divide, so we can take expectations of X and its powers. Then the q^{th} moment of X — for some reason it's always q — is just $\mathbf{E}X^q$.

3.3.1 Particularly Important Moments

Mean

The **mean** is the first moment, or simply the expectation value of X.

Variance and Standard Deviation

The second moment, $\mathbf{E}X^2$ doesn't have any particular name. The difference between the second moment and the square of the first moment — $\mathbf{E}X^2 - (\mathbf{E}X)^2$ — is called the **variance** of X, and is sometimes written $\mathrm{Var}X$.

The square-root of the variance is often called the **standard deviation**, which is often written σ.

Exercise. Convince yourself that $\mathbf{E}X^2 - (\mathbf{E}X)^2 = \mathbf{E}(X - \mathbf{E}X)^2$. Further convince yourself that this quantity is never less than zero.

Exercise. Convince yourself that if X and Y are independent, then the mean of their sum is the sum of their means, and that $\mathrm{Var}(X + Y) = \mathrm{Var}X + \mathrm{Var}Y$.

Chapter 4

Important Discrete Distributions

We're only going to look at two; there are about a half-dozen others which are common in practice, or at least in theory, and which you'll find discussed in any good textbook, such as Grimmett and Stirzaker (1992), or in the reference manuals, e.g. Patel, Kapadia and Owen (1976).

4.1 The Bernoulli Distribution

Here the space is very simple: $\{0, 1\}$, or any other two-element space: heads or tails, rain or shine, Democrat or Republican. The probability of getting a one is the parameter of the distribution, which is conventionally written either p or μ. That is, $\mathrm{P}(X = 1) = p$ or $= \mu$.

By itself, a Bernoulli variable isn't very interesting. But a string of independent Bernoulli variables with the same parameter is very interesting, because very powerful theorems can be proved about them fairly easily, so a lot of probabilistic effort goes into making things look like Bernoulli variables. In particular, think of your favorite event in your favorite probability space. Now either that event happens or it doesn't; so we write down a 1 when it does and a 0 when it doesn't, and presto, we have a Bernoulli random variable, and can apply all our theorems to it. We'll see shortly how this can be important.

4.2 The Binomial Distribution

We have a great number of objects which come in two sorts — the classical example, the gods alone knows why, is white and red marbles in an urn. A fraction p of them are of one sort (red, say), and we pick out N of them at random. We care about how many are red, but not about any sort of order or what-not might be among them. So our random variable, X, is the number of

red balls (or more generally **successes**). The distribution of X is the **binomial distribution**. It is also the distribution of a sum of N Bernoulli variables, and while deriving it is a fine way of building character, I'll just state it:

$$\mathrm{P}(X = x) \equiv \left(\begin{array}{c} N \\ x \end{array} \right) p^x (1-p)^{N-x} \tag{4.1}$$

The last two terms on the right-hand side are easy enough: they're the probability of getting a success x times and failing the other $N - x$ times, if every ball is independent of the others. But what's the ugly thing up front? It's read "from N choose x," or just "N choose x", and it's the number of ways of picking x objects from N objects, without replacement and without caring about the order. It's $= \frac{N!}{x!(N-x)!}$, where $x!$ is "x factorial".

4.3 The Poisson Distribution

The Poisson distribution is

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} , \tag{4.2}$$

for $k \geq 0$ and $\lambda > 0$. It is the limit of the binomial distribution as $N \to \infty$ and $p \to 0$ but $Np \to \lambda$ — the total expected number of successes remains fixed, even as their density goes to zero. We'll encounter this distribution again under stochastic processes.

4.4 First Moments of These Distributions

Distribution	$\mathrm{P}(X = x)$	$\mathbf{E}X$	$\mathrm{Var}X$	Comments
Bernoulli	$\begin{array}{ll} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{array}$	p	$p(1-p)$	$0 < p < 1$
Binomial	$\left(\begin{array}{c} N \\ x \end{array} \right) p^x (1-p)^{N-x}$	Np	$Np(1-p)$	N integer, $0 < p < 1$
Poisson	$\frac{\lambda^x}{x!} e^{-\lambda}$	λ	λ	$k \geq 0$, $\lambda > 0$

Chapter 5

Continuous Random Variables

The tricky thing about continuous random variables is that you can't make points your basic events. The problem is related to the fact that every point has length 0, but if you string enough points together you have something with positive length. The usual way around this difficulty, for us, is to define probabilities for intervals. (I'll just talk about one-dimensional continuous distributions; you need a bit more trickery for higher-dimensional distributions, but not too much more.)

5.1 The Cumulative Distribution Function

The standard trick is to consider intervals which stretch from $-\infty$ to your favorite point x, and have a function which gives you the probability of each of them. There's something of a convention to write that function with a capital letter. So $F(x) = \mathrm{P}(-\infty < X \leq x)$ gives us the probability that $X \leq x$. This is called the **cumulative distribution function** or **CDF** for X.

If we want the probability for some other interval, say from a to b, we just subtract: $\mathrm{P}(a \leq X \leq b) = F(b) - F(a)$.

Exercise. Convince yourself that this last statement is right, using the axioms for probabilities of events. Remember that disjoint intervals are mutually exclusive events, so their probabilities will add.

5.2 The Probability Density Function

If you use the CDF to find the probability of smaller and smaller intervals around your favorite point, you'll get smaller and smaller numbers. The natural thing for a mathematician to do is to divide those probabilities by the length of the intervals; then you've got a derivative, which tells you how much probability

there is in the neighborhood of your favorite point. This is the **probability density function**, or **pdf**, or probability density, density function, or even probability derivative function. More formally, $p(x) = \frac{dF(x)}{dx}$. The probability of a set A is now

$$\mathrm{P}(X \in A) = \int_{x \in A} p(x)dx \ . \tag{5.1}$$

Some people prefer to write dp instead of $p(x)dx$. (There are even cases where it is a Good Thing, but I won't get into that.)

Exercise. Convince yourself that this gives the same result for intervals as using the CDF.

Some otherwise-satisfactory CDFs do not have well-defined derivatives, so they don't have pdfs. This is a bitch, but not one we're likely to encounter here.

There are conditional CDFs and pdfs, just as you might expect. The existence of conditional pdfs is even trickier than that of regular pdfs, but again, we probably won't have to worry about that.

5.3 Continuous Expectations

The expectation of a function f of a continuous random variable X, with pdf $p(x)$, is simply

$$\mathbf{E}f(X) \equiv \int f(x)p(x)dx \ . \tag{5.2}$$

All the properties of expectations in the discrete case generalize, substituting integrals for sums where necessary. Conditional expectations, moments, etc., are all defined analogously.

Chapter 6

Important Continuous Distributions

Again, we're going to really limit ourselves here to some particularly important examples. Again, for more distributions and more details, see Patel, Kapadia and Owen (1976).

6.1 The Exponential Distribution

This is simply the result of a constant *rate* of decay, i.e., there is a constant probability per unit time λ of disappearing. decay. The CDF is $F(x) = 1 - e^{-\lambda x}$, so that the pdf is $p(x) = \lambda e^{-\lambda x}$. (Observe that this is the solution to $\frac{dp(x)}{dx} = -\lambda p(x)$.)

6.2 The Normal or Gaussian Distribution

So-called after Karl Friedrich Gauss, who did not discover it. This is the classic bell-curve.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.1}$$

This is the single most important distribution in probability theory, owing to the Central Limit Theorem. (See below). When we want to refer to a Gaussian distribution with mean μ and variance σ^2, we'll write $\mathcal{N}(\mu, \sigma^2)$.

If X $\mathcal{N}(\mu_1, \sigma_1^2)$ and Y $\mathcal{N}(\mu_2, \sigma_2^2)$, and X and Y are independent, then $aX + bY$ $\mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$. This will often be useful later on.

6.3 The χ^2 Distribution

This is very important in hypothesis testing and other parts of statistics. There is one parameter, an integer d. The pdf is

$$p(x) = \frac{1}{\Gamma(\frac{d}{2})2^{d/2}} x^{\frac{d}{2}-1} e^{-x/2} \tag{6.2}$$

where $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$. This has the nice property that $\Gamma(t+1) = t\Gamma(t)$, and in fact for integer t, $\Gamma(t) = (t-1)!$. Since d is an integer, the only other fact we need to compute it for our purposes is that $\Gamma(1/2) = \sqrt{\pi}$.

The parameter d is called the number of **degrees of freedom**. We write this distribution as $\chi^2(d)$. The importance of this distribution comes from the fact that the square of a normal random variable is distributed as $\chi^2(1)$, and from the fact that if X and Y are independent, with distributions $\chi^2(n)$ and $chi^2(m)$, then $X+Y$ is $\chi^2(n+m)$.

6.4 The Lognormal Distribution

So-called for obvious reasons:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} \tag{6.3}$$

That is to say, the logarithm of x has a Gaussian distribution. Thus here μ and σ are the mean and standard deviation of the logarithm of X, not of X itself.

Note that $\mathbf{E}\log x = \mu$, but that doesn't mean that $\log \mathbf{E}x = \mu$! (See sec. 6.6.)

6.5 Power-Law Distributions

The pdf for this distribution is

$$p(x) = kx^{-\alpha}\ , \tag{6.4}$$

where $\alpha > 0$. Note that x cannot go through 0, since then the integral which ought to give us the CDF diverges. So power-laws describe parts of distributions, those above (or below) a certain cut-off value. The simplest way of doing this is to simply say that we never observe values below the cut-off, giving us what's called the **Pareto distribution**,

$$F(x) = 1 - \left(\frac{a}{x}\right)^{\alpha-1} \tag{6.5}$$

when $x \geq a$, and $\mathrm{P}(X < a) = 0$.

People around SFI are very into power-law distributions, for reasons which probably will become clear over the course of the summer school.

6.6 First Moments and pdfs of These Distributions

Distribution	pdf	$\mathbf{E}X$	Var$X$	Comment
Exponential	$\lambda e^{-\lambda x}$	λ^{-1}	λ^{-2}	$\lambda > 0$, $x \geq 0$
Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	
χ^2	$\frac{1}{\Gamma(\frac{d}{2})2^{d/2}} x^{\frac{d}{2}-1} e^{-x/2}$	d	$2d$	d integer
Lognormal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}$	$e^{\mu+\frac{\sigma^2}{2}}$	$e^{\sigma^2}\left(e^{\sigma^2}-1\right)e^{2\mu}$	$x > 0$
Power Law (Pareto)	$\frac{\alpha a^\alpha}{x^{-(\alpha+1)}}$	$\frac{\alpha a}{\alpha-1}$ if $\alpha > 1$	$\frac{\alpha a^2}{(\alpha-1)^2(\alpha-2)}$ if $\alpha > 2$	$\alpha > 0$, $x \geq a$

Chapter 7

Tricks with Random Variables

In the next two sections, $X_1, X_2 \ldots X_n \ldots$ are independent, identically-distributed (IID) random variables. X has the finite mean μ. The average of the first n of them is

$$S_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \; . \tag{7.1}$$

7.1 The Law of Large Numbers

Theorem. The means of IID sequences converge to the mean of the variables:

$$S_n \xrightarrow[n\to\infty]{} \mu \; . \tag{7.2}$$

This will probably satisfy most of you in most applications. At other times, you ought to worry about just what it is I mean by "converge".

The weakest sense is that the cumulative distribution functions of the S_n will converge on the CDF of the "random variable" which is always μ. (That CDF is 0 if $x < \mu$, and 1 if $x \geq \mu$.) This is called "convergence in distribution", and requires merely that $\mathbf{E}X$ is well-defined and finite.

The strongest sort of convergence, "almost-sure" convergence, is when, for almost all realizations of the random variables X_i, $S_n \to \mu$, i.e., when the convergence happens with probability one. The necessary and sufficient condition for this is that $\mathbf{E}|X| < \infty$. (It may be easier to show that $\mathbf{E}X^2 < \infty$, which implies a finite expectation for the absolute value.) Such variables are said to satisfy the "strong law of large numbers". [1]

[1] Just to make matters confusing, the "weak law of large numbers" is *not* converge-in-distribution to the mean. Rather, it is convergence in probability: $\mathrm{P}(|S_n - \mu| > \epsilon) \xrightarrow[n\to\infty]{} 0$ for any positive ϵ. The necessary and sufficient condition for this is a pair of ugly integrals I won't bother you with (Grimmett and Stirzaker 1992, ch. 7).

No form of the law of large numbers says anything about how *fast* the S_n converge to μ. What's wanted here is a result which says something like $\mathrm{P}(|S_n - \mu| > \epsilon) \leq f(n, \epsilon)$, for some function f which goes to zero as n goes to infinity. For IID variables, the general result, called the **large deviation principle** (den Hollander 2000), is that, as $n \to \infty$, $\mathrm{P}(|S_n - \mu| > \epsilon) \to e^{-ng(\epsilon)}$, where $g(\epsilon)$ is some polynomial, called the **rate function**[2]. But the large deviation principle is another asymptotic result; there *are* rate of convergence results which are valid at all n, but the generally depend on the kind of distribution we're looking at. For more on this, and the applications of rate of convergence results to statistics, see Vapnik (2000) (highly recommended), Pronzato, Wynn and Zhigljavsky (1999) and van de Geer (2000).

The obvious application of the law of large numbers is to take our favorite event A from any probability distribution we like, with random variable Y, and then apply an **indicator function** I_{A} to Y, a function which gives 1 when its input is in in A and 0 otherwise. Now we have a Bernoulli random variable with parameter $p = \mathrm{P}(\mathrm{A})$. This has finite mean and variance, so it has a finite second moment, and consequently the strong law applies. So if we take a long sequence of random trials, then with probability one the frequency of successes will converge to the true probability (and by the large deviation principle, they'll converge exponentially fast).

7.2 The Central Limit Theorem

Theorem. Assume that X has a finite variance σ^2. Then as $n \to \infty$, the S_n converge to a Gaussian random variable with mean μ and variance σ^2.

Note 1. The convergence here is convergence-in-distribution (see previous section).

Note 2. The conditions needed for the density function to converge are a bit stronger (Grimmett and Stirzaker 1992, sec. 5.10).

Note 3. Note again that the theorem doesn't say anything about the rate of convergence. This can again be a tricky subject, though an admirably simple result is available if the X_i are themselves Gaussian...

7.2.1 The Extraordinary Importance of the CLT

Thus Francis Galton in 1886:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by 'the law of error.' A savage, if could understand it, would worship it as a god. It reigns with severity in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the anarchy the more

[2] More exactly, $\lim_{n\to\infty} \frac{1}{n} \log \mathrm{P}(|S_n - \mu| > \epsilon) = g(\epsilon)$. The problem with the statement in the main text is that, since the probability goes to zero as n goes to infinity, it will converge on *any* function which does likewise, and we want to pick out the *exponential* convergence.

> perfect its sway. Let a large sample of chaotic elements be taken and marshalled in order of their magnitudes, and then, however wildly irregular they appeared, an unexpected and most beautiful form of regularity proves to have been present all along. [3]

More prosaicly: the CLT tells us that, if only we can arrange to be dealing with IID sequences, in the long run we know what the distribution *must* be. Even better, it's always the same distribution; still better, it's one which is remarkably easy to deal with, and for which we have a huge amount of theory. Manipulate your problem so that the CLT applies, and, at least asymptotically, you've got it made.

7.2.2 Ways in Which the CLT Can Fail

First: the variables may not be independent. There are some situations in which this can be overcome (Manoukian 1986).

Second: the variables may not have well-defined variances. There are quite respectable probability distributions with a well-defined mean, but where the integral for the variance diverges. (Some of these, important in physical theory, are called "Lévy walks".) No variance, obviously no sequence of variables with equal variance!

Third: the variables might not be identically-distributed. There are some generalizations here, too, supposing at least that the means are identical (Manoukian 1986).

7.2.3 The Many-Independent-Causes Story for Why Things Are Gaussian

It is often asserted that in many situations errors in measurement are at least roughly Gaussian. As someone who taught introductory physics for four years, I can assure you that this is actually true. There is even a story for why this should be so. Think of your favorite measurement activity — let us say, measuring the height of a pendulum-string with a meter-stick. There are lots of things which can make this go wrong: the stick may bend or warp; the markings may be smudged, slanted or twisted; the string may bend; its ends may be frayed; it may move; the eye will not be exact; I could go on. Now suppose that the effects of each of these causes simply sum up to give the total error; that the mean error induced by each cause is 0 (i.e., that they introduce no *systematic* bias into measurement); and that the causes are statistically independent. Then we should expect that the sum — the total error — will have a roughly Gaussian

[3] As the historian and philosopher Ian Hacking notes, on further consideration Galton was even *more* impressed by the central limit theorem, and accordingly replaced the sentence about savages with "The law would have been personified by the Greeks and deified, if they had known of it." If it had been discovered in our time, I daresay we would have called it "self-organized normality," and written books about it with titles like *How Nature Works*.

distribution, and the more independent causes, the better the fit to a Gaussian.[4]

Generalizing this story, whenever we have additive effects of many independent causes, each of which has only a small relative effect, we are inclined to suspect Gaussians. Thus the many biological examples of Gaussian distributions (e.g., chest widths of Scots army recruits, annual number of sucides in Belgium) which so impressed nineteenth century authors like Galton.

7.2.4 Multiplicative Noise and the Lognormal Distribution

Suppose that we have variables X_i which, as before, are IID, and have a finite mean and variance, but that instead of adding them together and dividing, we multiply and take the n^{th} root ("the geometric mean"), i.e.,

$$S_n = \left(\prod_{i=1}^{n} X_i\right)^{1/n} \text{ so that} \tag{7.3}$$

$$\log S_n = \frac{1}{n}\sum_{i=1}^{n} \log X_i \ . \tag{7.4}$$

Then, assuming that $\log X_i$ has finite mean and variance, it follows that $\log S_n$ will be normally distributed, and the distribution of S_n will be log-normal. So if the X_i represent noise variables which multiply instead of adding, and their logarithms are well-behaved, then the effect of lots of small sources of noise will be to give us a lognormal distribution for the total noise.

Notice that it's often very hard to distinguish lognormal distributions from power-law distributions, especially when over-enthusiastic people forget to check for this possibility! This is a primary reason why you should be suspicious of claims that such-and-such a process produces power-laws.

Aside. Let's expand on that last part just a little. Recall that the pdf for a power-law is $p(x) = kx^{-\alpha}$. So $\log p = logk - \alpha \log x$, and if we make a log-log plot of the density curve we get a straight line; very often when people claim to have found a power-law, they mean nothing more than that they got a decent fit to a straight line on such a curve. Now let's do the same thing for the lognormal pdf.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} \tag{7.5}$$

$$\log p = -\log\sqrt{2\pi\sigma^2} - \frac{(\log x - \mu)^2}{2\sigma^2} \tag{7.6}$$

$$\log p = b - \frac{\mu}{\sigma^2}\log x - \frac{(\log x)^2}{2\sigma^2} \ , \tag{7.7}$$

[4] You may be worrying about the fact that the causes of errors are independent but not identically distributed. There are theorems which say that even in this case the sum should be Gaussian, provided some conditions on moments are met.

where we've lumped all the constants into one term b. If $x \approx 1$, then $\log x$ is small and $(\log x)^2$ is negligible, so

$$\log p \approx b - \frac{\mu}{\sigma^2} \log x \tag{7.8}$$

which, of course, looks like a power-law with exponent $\alpha = \frac{\mu}{\sigma^2}$.

Chapter 8

The Care and Handling of Data

> Sit down before fact as a little child, be prepared to give up every preconceived notion, follow humbly wherever and to whatever abysses Nature leads, or you shall learn nothing.
> — T. H. Huxley, letter to Charles Kinsley, 23 September 1860

> Data simply *occur* to me.
> — Dr. Science

Dr. Science is lucky; most of us have to work for our data, earning it with the sweat of our instrumentation; even theorists like me have, these days, near-data in the form of simulation results. Data are accordingly quite precious, and, as Huxley indicates, they are not to be lightly tampered with. We'll start looking at statistics by looking at ways of summarizing or describing data which don't rely on any sort of auxiliary hypothesis. (I am going to assume throughout that the data are numbers.)

8.1 Counting

First, some ways of dealing with data which involve mere counting.

8.1.1 The Question of Bins

Before continuous data can be counted up, they need to be divided into discrete blocks, or *binned*. A natural question is, what sort of bins to use? This is tricky; generally you want the bins to be large and coarse enough that there's a reasonable degree of representation in each, but not so coarse that you lose all structure. If you make the bins too small, then the variance in the expected number of points in each bin will get too big.

It's generally a bad idea to have variable-sized bins, but not always.

It may be a good idea to experiment with different bin-sizes until you find a regime where the bin-size doesn't have much effect on your analysis. This unfortunately is quite time-consuming.

8.1.2 Histograms

A histogram is simply a plot of how many data-points fall into each bin. As such, it's a kind of approximation to the probability density function. (There are actually quite sophisticated techniques for estimating the pdf from histogram data.)

Some people want to use the word "histogram" only for one-dimensional data; I don't see why.

8.1.3 Percentiles

Assuming the data are one-dimensional, so they can be put in a simple order, then we can talk about the percentiles — just like on the GREs. The x^{th} percentile value is simply the value equaled or exceeded by only $\frac{x}{100}$ of the data-points. That is, x percent of the data are at or above that value. Similarly for deciles and quartiles.

Just like the histogram is an approximation to the pdf, the percentiles contain information about the cumulative distribution function.

8.1.4 Median

Again assuming one-dimensional data, the median is the value such that half of all points have a higher value and half a lower. If there is a gap in the data, there may well be a range of values which can claim to be the median; there is a weak convention, in such cases, to use the mid-point of the range.

8.1.5 Mode

The mode is simply the value with the most data-points.

8.2 Adding

8.2.1 Sample Mean

The sample mean, $\hat{\mu}$, is just what you'd expect:

$$\hat{\mu} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \ . \tag{8.1}$$

8.2.2 Sample Variance

The sample variance, s^2, is again what you'd expect:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \; . \tag{8.2}$$

Now here comes a subtlety: the sample variance is *not* a good estimate of the population variance; it's too small. The population variance is best estimated by $\frac{n}{n-1}s^2$. (That's why I write s^2 and not $\hat{\sigma}^2$.) The story here, heuristically, is that you tend to lose variation under sampling, so measures of variation in the sample need to be corrected upwards. A slightly more sophisticated story is that, if samples are independent, some analysis of the sampling distribution shows that $\mathbf{E}S^2 = \frac{n-1}{n}\sigma^2$, so this is the right correction to use. A still more sophisticated story claims that what's really important in estimation, and what we really should divide through by, is not the number of data-points but the number of "degrees of freedom," and that to get the variance we need to estimate the mean, thereby losing one degree of freedom.

The bottom line is that, while you should use $\frac{n}{n-1}s^2$ as your estimate of the population variance, if the difference between that and s^2 is big enough to matter, you probably should think about getting more data points!

8.3 Correlating

"Correlation" is used in a specific sense in statistics; it almost always refers to a linear relationship (plus noise) between two random variables: when one goes up, the other goes up, on the average by an amount proportional to the increase in the first variable.

8.3.1 Covariance

One way of measuring correlation between two variables is their **covariance**:

$$\mathrm{Cov}(A, B) \equiv \mathbf{E}AB - (\mathbf{E}A)(\mathbf{E}B) \tag{8.3}$$

Exercise. Convince yourself that this should be 0 when A and B are independent.

Covariations are often used in statistical physics, where for some reason we call them **correlation functions**.

8.3.2 The Correlation Coefficient

a.k.a. Pearson's correlation coefficient, is just a normalized covariance:

$$\mathrm{Corr}(A, B) \equiv \frac{\mathrm{Cov}(A, B)}{\sqrt{(\mathrm{Var}A)(\mathrm{Var}B)}} \tag{8.4}$$

The correlation coefficient is thus constrained to lie between -1 and $+1$ inclusive. The extreme values indicate perfect linear dependence.

Exercise. Convince yourself of this.

Instead of writing $\mathrm{Corr}(A, B)$, some people write ρ_{AB}, or even just ρ.

Chapter 9

Sampling

We assume that whatever it is that gave us our data is well-modeled by some random process. In effect, each time we make a measurement, we ask the process to spit out numbers according to some distribution; we **sample** the distribution. If we think of sampling as closing our eyes and picking a data point out of a box, we call the set of all data-points in the box the **population**. (That last isn't as silly as it sounds; think of opinion polls.) In either case, we assume that the *true* distribution is the unknown one which lies behind our data. If we want to learn about that distribution from the data, then we need to know something about the kind of data it is likely to give us. That is to say, we need to know the **sampling distribution**, the distribution of data values given that the true distribution takes a certain functional form, with certain parameter values, and we sample from it in a specified way.

9.1 The Notion of "A Statistic"

We want to ignore as much about our data as we can get away with; we want to summarize it, rather than having to remember (say) the exact sequence of heads and tails for a million coin-tosses. A summary of the data is called a "statistic" in the trade. More formally, any function of the data is a statistic, provided (1) it's well-defined for any number of data-points, (2) it has no random inputs other than the data. We include constants functions as degenerate cases.

9.2 Loss of Variability Under Sampling

It's generally true that, whatever measure of variability we pick, its value in a sample will tend to be smaller than its population value. (We have already mentioned that this is true of the variance.) It's not hard to see why this should be so; consider, as an extreme case, the limit where our sample consists of a single piece of data, and so there is no variation in the sample. More generally, picking out a subset from the population, which is what sampling

does, is unlikely to give an exactly representative sample, and naturally the higher probability events will be the ones which tend to show up in the sample.

Exercise. Consider a Bernoulli variable with $p = 0.01$. If we sample this 100 times, what is the probability of not getting *any* successes in our sample?

Loss of variability due to sampling is important in evolutionary genetics, since it is the origin of genetic drift (Gillespie 1998).

9.3 Figuring the Sample Distribution; Monte Carlo Methods

Even if you know what the exact population distribution is, it can be extremely hard to figure out what the sampling distribution is. In such cases you have three alternatives:

- Do the work yourself.
- Turn to the literature in the hopes that somebody else has already done it.
- Simulate.

The virtues and the drawbacks of the first two courses of action speak for themselves. The third is a bit trickier. If you know what your sampling procedure is, and you know what the population distribution is, and you have a good source of random numbers, then in principle you can simulate a sample from the population. If you simulate *many* samples, then the laws of large numbers tell us the the empirical frequencies of your simulations will approach the actual sampling distribution. This is called "Monte Carlo simulation", or even "the Monte Carlo method". (Several other things are also called Monte Carlo, but they're actually related.) By the same procedure, you can get the sampling distribution for an arbitrary statistic of your data.

The obvious question, then, is how many times you have to run your simulation before you can be confident that you're pretty close to the right distribution. *That* is tricky. Sometimes a few hundred points is enough; sometimes you need hundreds of thousands or more. For most classroom-type problems, you can get away with using a few thousand or ten thousand points, but I recommend looking at what *Numerical Recipes* (Press, Teukolsky, Vetterling and Flannery 1992a; Press, Teukolsky, Vetterling and Flannery 1992b) has to say, and, if it's really important, books like Mark Newman's treatise on Monte Carlo methods (Newman and Barkema 1999), or MacKeown (1997).

Chapter 10

Estimation

10.1 Point Estimates

A **point estimate** is a statistic which, given the data and an assumption that the distribution is of a certain form, gives us a guess at to what one of the distribution's parameters is. The statistic (or the function it applies to the data) is called an **estimator**. We say it's a *point* estimate because it only returns a single value for the parameter; other kinds of estimates give us ranges.

Conventionally, the possible values of the parameter are θ, the true value is θ_0, and the estimator is $\hat{\Theta}$ (for the random variable) or $\hat{\theta}$ (for its particular value given our data).

10.1.1 Bias, Variance, and Other Sorts of Quality

An estimator $\hat{\Theta}$ is **consistent** if $\hat{\Theta}$ converges to the true parameter value θ_0 as the number of data-points n goes to infinity.[1]

The **bias** of an estimator $\hat{\Theta}$ is the expected error in its estimate: $\mathbf{E}\hat{\Theta} - \theta_0$. In general, the bias is a function of θ_0 (since $\mathbf{E}\hat{\Theta}$ is). If the bias is zero whatever the true value of the parameter, then $\hat{\Theta}$ is **unbiased**. Note that a consistent estimator is not necessarily unbiased; the sample variance is a consistent estimator of the population variance, but it has a negative bias. (See section 8.2.2.)

Exercise. Convince yourself that the sample variance is a consistent estimator of the true variance.

The variance of $\hat{\Theta}$ is simply its normally defined variance. In general, the variance depends on the number of data-points, and should go down as the number of data-points goes up.

The mean square error of $\hat{\Theta}$ is $\mathbf{E}(\hat{\Theta} - \theta_0)^2 = \left(\mathbf{E}\hat{\Theta} - \theta_0\right)^2 + \mathrm{Var}\hat{\Theta}$. That is, the mean square error is the bias squared plus the variance.

Exercise. Convince yourself of this.

[1] Technically, converges in probability.

We want our estimators to be unbiased and to have very small variance. For a fixed number of data-points n and the same bias, we prefer the estimator which has the lower variance. One can establish a lower bound on the mean square error which depends only on the bias, n, and the population distribution (the **Cramér-Rao inequality** or **information inequality**; see Cramér (1945, sec. 32.3)). In the case of unbiased estimators, this establishes a lower bound on the variance of the estimate. An unbiased estimator which has the minimum variance for any n is called **efficient**. An estimator whose variance approaches the minimum as $n \to \infty$ is **asymptotically efficient**.[2]

A **sufficient statistic** for a parameter is one where the distribution of samples, conditional on the statistic, is independent of the parameter. That is, T is a sufficient statistic for θ iff $p_\theta(x|T=t)$ is independent of θ. In this sense, T summarizes all the information the data contain about the value of θ. As a consequence, for any function f of the data x, $p_\theta(f(x)|T=t)$ is also independent of θ. As another consequence (the **Neyman factorization theorem**), there are functions g and h such that $p_\theta(x) = g(\theta, t)h(x)$, where h has no dependence at all on θ. A sufficient statistic T is **minimal** if it is a function of every other sufficient statistic R, i.e. if for every sufficient R there is an f such that $T = f(R)$.

There are two reasons by sufficiency statistics are important. First, they lead to lower-variance estimates (the **Rao-Blackwell theorem**). Say D is an unbiased estimator of θ. Then one can show (Lehmann and Casella 1998, theorem 1.7.8) that $\delta(t) = \mathbf{E}(D(X)|T=t)$ is also an unbiased estimator of θ, and that $\mathrm{Var}\delta < \mathrm{Var}D$, unless $D(X) = \eta(T(x))$ with probability one, in which case the variances are equal. Second, sufficiency is invariant under a change of coordinates in the parameters. If T is sufficient for θ, then for any nice function h, $h(T)$ is sufficient for $h(\theta)$.

Every efficient estimator is a sufficient statistic for its parameter.

(All this is about **parametric sufficiency**. There is also the notion of **predictive sufficiency**, where $T(X)$ is sufficient for predicting Y from X if $\mathrm{P}(Y|X) = \mathrm{P}(Y|T(X))$. It has similar desirable properties.)

10.1.2 Some Common Kinds of Estimates

Least Squares

It is very common to measure errors by the square of the difference between what is predicted by a hypothesis and what is actually measured. So, for instance, if

[2] There are two tangent directions here.

One is that, given that there's a fixed lower bound on the error, which is the sum of the bias and the variance, we may sometimes be able to decrease one of these only by increasing the other. This is sometimes called the **bias-variance tradeoff** in non-parametric estimation and machine learning (Zapranis and Refenes 1999).

The other tangent is that the kind of reasoning employed in the proving the Cramér-Rao inequality can be generalized, leading to connections between estimation error and a quantity known as the Fisher information matrix, which in turn connects to information theory (Kullback 1968) and the design of experiments (Atkinson and Donev 1992).

we expect that a curve will follow $y = f(x, \theta)$, and we measure n pairs of values (x_i, y_i), the mean square error is

$$E^2(\theta) = \frac{1}{n} \sum_{1}^{n} (y_i - f(x_i, \theta))^2 \ . \tag{10.1}$$

The **least squares estimate of** θ is

$$\hat{\theta}_{LS}(x_1, \ldots x_i, \ldots x_n) = \arg\min_{\theta} E^2(\theta) \ . \tag{10.2}$$

Notice that, if $f(x, \theta)$ is a probability distribution, and the y are observed frequencies, so that (x_i, y_i) is how often we observe the event x_i, we can use least squares to estimate the distribution's parameters.

Note that least squares, in this form, takes no account of how much variance our hypothesis says there *ought* to be at a given value of x. A modification which does this is minimizing χ^2 — see sec. 11.1.2 — which is also more robust, in general, than least squares.

Maximum Likelihood

The **likelihood** of getting data $x_1, x_2, \ldots x_n$ is simply the probability of seeing all those values of the data, given a certain value of the parameter : $L(\theta|x_1, x_2, \ldots x_n) = \mathrm{P}_\theta(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$. Taking the data as fixed, we ask what value of the parameter maximizes the likelihood:

$$\hat{\theta}_{ML}(x_1, \ldots x_i, \ldots x_n) = \arg\max_{\theta} L(\theta|x_1, x_2, \ldots x_n) \ . \tag{10.3}$$

It is important to keep straight what this means. The maximum-likelihood value of θ is the one which makes our data as probable as possible; it is *not* the most probable value of the parameter given our data.

Notice that maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood. If the data are independent samples,

$$L(\theta|x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} L(\theta|x_i) \tag{10.4}$$

$$\log L(\theta|x_1, x_2, \ldots x_n) = \sum_{i=1}^{n} \log L(\theta|x_i) \tag{10.5}$$

$$= \sum_{i=1}^{n} \log \mathrm{P}_\theta(x_i) \tag{10.6}$$

and finding the θ which maximizes *that* is much easier than the general problem.

Maximum likelihood estimates are extremely common, because, under some fairly mild conditions which I shan't spell out (Lehmann and Casella 1998), they are guaranteed to be both consistent and to have a Gaussian distribution about the true parameter θ_0, at least in the limit where $n \to \infty$. In the case of Gaussian errors, least-squares estimates and maximum-likelihood estimates coincide, but ML estimation is much more widely applicable than least-squares is.

10.2 Curve-Fitting or Regression

Suppose we have data consisting of pairs of input values (**independent variables**) x and output values (**dependent variables**) y. We assume that there's some kind of functional relationship between the input and the output, so that $Y = f(x, \theta) + \eta$, where η is some kind of random noise. Assuming the functional form is fixed, and the character of η is known, **regression** is the problem of making a good guess at θ. Technically, this is a sort of parameter estimation problem, since θ is one of the parameters for the distribution of Y at each value of x. (The other parameters are in the noise term.) In the case of **linear regression**, we assume that $Y = a + bx + \eta$, and generally that η is a mean-zero Gaussian, independent for each measurement, and then we ask for good values of a and b. Linear regression involves little more than linear algebra, but it's a bit too involved for me to go into here; see, among many other books, Gonick and Smith (1993) and Weisberg (1985).

Both input and output values can be vectors, not just scalars.

10.3 Propagation of Errors

We often wish to calculate quantities which are functions of other, measured quantities. Given that there is some degree of uncertainty in the measurements, there will be uncertainty in the value of the quantities that depend on them. If you want, you can regard these calculated quantities as statistics of a sort, and try forming point estimates and confidence intervals for them. A much cruder but often adequate way of dealing with these uncertainties is what's called **propagation of errors**, which rests on the assumption that errors of measurement are Gaussian.

Consider a quantity $z = f(x, y)$. We know the arguments only imperfectly, so we model them as independent random variables, $X \sim \mathcal{N}(\bar{x}, \sigma_x^2)$ and $Y \sim \mathcal{N}(\bar{y}, \sigma_y^2)$. Together they give us a new random variable $Z = f(X, Y)$. What is Z like?

Let us apply Taylor's Theorem:

$$\begin{aligned} Z &= f(\bar{x}, \bar{y}) \\ &\quad + \left(\left. \frac{\partial f}{\partial x} \right|_{x=\bar{x}} \right) (X - \bar{x}) \\ &\quad + \left(\left. \frac{\partial f}{\partial y} \right|_{y=\bar{y}} \right) (Y - \bar{y}) + \text{higher order terms} \end{aligned} \tag{10.7}$$

We assume that higher order terms are negligble. Now, $X - \bar{x}$ and $Y - \bar{y}$ are independent Gaussians with mean 0, and this is not changed by multiplying them by the partial derivatives in front. (Call those partials f_x and f_y, for convenience.) And if we add independent Gaussians with mean 0, we get another Gaussian with mean 0, so Z is going to be a Gaussian with mean $\bar{z} = f(\bar{x}, \bar{y})$.

What, however, is the variance of Z?

$$\begin{aligned} Z - \bar{z} &= f_x(X - \bar{x}) + f_y(Y - \bar{y}) && (10.8) \\ (Z - \bar{z})^2 &= f_x^2(X - \bar{x})^2 + f_y^2(Y - \bar{y})^2 + 2f_x f_y(X - \bar{x})(Y - \bar{y}) && (10.9) \\ \mathbf{E}(Z - \bar{z})^2 &= f_x^2 \mathbf{E}(X - \bar{x})^2 + f_y^2 \mathbf{E}(Y - \bar{y})^2 && (10.10) \\ \mathrm{Var}Z &= f_x^2 \sigma_x^2 + f_y^2 \sigma_y^2 && (10.11) \end{aligned}$$

So, in the end, $Z \sim \mathcal{N}(f(\bar{x}, \bar{y}), {f_x}^2 \sigma_x^2 + {f_y}^2 \sigma_y^2)$, and we have a standard deviation for z.

This extends to any number of variables, naturally.

10.4 Confidence Regions

One of the problems with point estimates is that they give us a single value. This may in some sense be the *best* single guess we could come up with for the value of the parameter, but it would be nice if we could get some idea of the *range* of good values — the range of what's reasonable, given our data. This is accomplished through a marvelously counter-intuitive construction due to Jerzy Neyman (Reid 1982; Cramér 1945), called **confidence intervals** or **confidence regions**.

The construction goes like this. We choose a statistic, call it X. Next we pick a probability α — one large enough that we don't mind a $1 - \alpha$ chance of being wrong. For each value of the parameter θ, we calculate an interval $C(\theta)$ such that $\mathrm{P}(X \in C(\theta)) = \alpha$. Generally we choose the intervals so that they're symmetric about $\mathbf{E}_\theta X$, but we don't always do this; we almost always choose them so that the end-points of the intervals are continuous functions of θ. We now perform our experiments and get a certain value x for the statistic. The confidence region is the set of all values of θ for which x falls within $C(\theta)$:

$$R(x) \equiv \{\theta | x \in C(\theta)\} \ . \quad (10.12)$$

(More visually, imagine a graph where θ runs along the horizontal axis and x runs along the vertical axis. Plotting $C(\theta)$ gives us two curves, such that the probability of X being between them is just α. Now we take a particular x and run a horizontal line across the graph at that height. It will (generally) cross each curve once, marking off an interval of θ. That is the α confidence interval compatible with x.)

Say we do this with $\alpha = 0.999$, and get a particular confidence region R for θ. This does *not* mean that there is a 99.9% chance that the true value θ_0 is in R — it either is or it isn't, and probability doesn't enter into it. What it does mean is that *either* $\theta_0 \in R$, *or* an unusual event, one with probability at most $1 - \alpha$, happened in our experiment. If we repeat the experiment many times, then R (a random variable, determined by the statistic X) will include θ_0 in α of the experiments.

So the meaning of the confidence region is "either the real parameter is in here, or we're very unlucky". Put this way, it becomes plausible — and is even true — that there is generally a trade-off between getting a tight estimate, having a small confidence region, and covering our asses. In the words of that eminent investigator C. Chan, "Improbable events permit themselves the luxury of occuring."

"Confidence interval" is generally used when θ is one-dimensional; "confidence region" is more general.

Formulæ for confidence regions for different standard distributions and sampling regimes are available in many textbooks and reference manuals. In the all-too-likely case that your sampling distribution is not among these, you can always construct the $C(\theta)$ by numerical integration or Monte Carlo, and then invert as normal.

Chapter 11

Hypothesis Testing

The basic idea is to ask, does this hypothesis fit the data better than the alternatives? We need to also ask, though, whether the fit is so much better that it couldn't, plausibly, be due to chance.

Obviously, the first thing we need is a way of measuring the fit between data and hypotheses.

11.1 Goodness-of-Fit

Traditionally, statistics which measure how well data accord with a statistical hypothesis are called **goodness-of-fit** measures. The name is a little misleading, because in almost all cases high values of these statistics indicate a very *bad* fit.

We have actually already looked at two goodness-of-fit statistics: the mean squared error and the likelihood. (High values of the likelihood actually indicate good fit.) Another and very important measure, however, is the χ^2 statistic, which is related to the mean square error, but more flexible. It has the advantage of letting us see whether or not the lack of fit is **significant**, an idea which I'll explain before getting to the test itself.

11.1.1 Significant Lack of Fit

We have a statistical hypothesis, which tells us that our data come from some distribution, with a parameter θ_0. Assume for simplicity that, if our hypothesis matches the data exactly, our goodness-of-fit statistic $\hat{G}$ will be 0, and that larger values of $\hat{G}$ become less likely, i.e., $\mathrm{P}_{\theta_0}(\hat{G} \geq g)$ is a monotonically-decreasing function of g. (There's no particular difficulty if $\hat{G}$ is not centered at 0, or if it can be either positive or negative.) We take our measurements and compute the value of $\hat{g}$. We now ask, what is the probability p that $\hat{G} \geq \hat{g}$, assuming the hypothesis is true? This is the **p-value** of the lack of fit.

An actual *test* of the hypothesis involves setting a threshold α, the **significance level** before taking the data, and rejecting the hypothesis if $p \leq \alpha$. That is, we reject the hypothesis if the deviation between it is highly significant, i.e., if it was very unlikely (had probability of at most α) of occuring by sheer chance.

11.1.2 The χ^2 Test

Suppose we have n data-points, $S_1, \ldots S_n$, and we want to test whether or not they conform to a certain hypothesis H_0, for which the cumulative distribution is $F(s)$. To do a χ^2 test, we first partition the range of S into m bins I_i. These bins do not have to be of equal size, but it's generally a good rule to have at least, say, five data-points in each bin. Then we say that p_i is the probability of falling into the i^{th} bin, if H_0 is true, i.e., $p_i = \mathrm{P}_{H_0}(S \in I_i)$. Similarly, we say N_i is the number of data points falling into the i^{th} bin. Now if H_0 is true, the expectation value of N_i is just np_i (remember we have n data-points). So we want to see if the deviation from expectation is too big. Now a bit of manipulation of binomial variables tells us that the deviation should be normally distributed for each bin, and we remember (sec. 6.3) that sums of squares of normals have the χ^2 distribution. So we make our test statistics

$$X^2 = \sum_{i=1}^{m} \frac{(N_i - np_i)^2}{np_i} \; . \tag{11.1}$$

You might expect, from what you know about the binomial distribution, that the denominator should be $np_i(1-p_i)$, but a careful analysis (Cramér 1945, sec. 30.1) gives us 11.1.

If H_0 is true, X^2 should be distributed like $\chi^2(m-1)$, at least in the limit of large n. We say that we have $m-1$ degrees of freedom in the data. Since the CDF of the χ^2 distribution is readily calculated and has often been tabulated, we can then see whether or not X^2 is so large that we can reject H_0 with a specified confidence level.

The above assumes that the same hypothesis, and so the same CDF, would have been used regardless of the data S_i. If instead we estimated r of the parameters from the data, we imposed r constraints and eliminated r degrees of freedom, and the distribution of X^2 should be $\chi^2(m-r-1)$.[1]

One of the most important applications of the χ^2 test is to regression problems. Suppose that we believe our dependent variable Y is a function of our independent variable x in such a way that $Y(x) = f(x,\theta) + \epsilon$, where θ is some parameter or set of parameters to a known functional form, and ϵ is a Gaussianly-distributed noise, with mean 0 and a standard deviation which depends, possibly, on x. Suppose we have m measurements, i.e., m pairs (x_i, y_i). For each pair, the error or residual is $y_i - f(x_i, \theta)$, and we expect these to have a Gaussian distribution, with mean 0 and standard deviation σ_i. Now if we look

[1] This assumes that the parameters are estimated in a "nice" way. For details, see Cramér (1945, sec. 30.3). The "-1", incidentally, is because the data are always under the constraint that the N_i sum up to n, i.e., there is always at least one constraint.

at $(y_i - f(x_i, \theta))^2/\sigma_i^2$, we expect that to be the square of a Gaussian variable with mean 0 and variance 1, so it should have the distribution $\chi^2(1)$. And if we add these squared, normalized residuals up over all our pairs,

$$X^2 = \sum_{i=1}^{m} \frac{(y_i - f(x_i, \theta))^2}{{\sigma_i}^2} \tag{11.2}$$

we should get something which is distributed like $\chi^2(m)$. Note that all of this is true even if some of our m data-points have the same value of x_i, i.e., we make more than one measurement with the same value of the independent variable. (This assumes that errors are independent between measurements.) Again, if we estimated r parameters from the data — that is to say, r components of θ — the number of degrees of freedom needs to be reduced by r, and so we need to look at the distribution of $\chi^2(m - r)$.

11.2 The Null Hypothesis and Its Rivals

In the goodness-of-fit cases we've just talked about, we had a hypothesis in mind, and tried to see if it matched the data. This baseline hypothesis is more formally called the **null**, and all standard hypothesis testing consists of asking whether the data force us to reject it, in favor of one of our alternatives. (If we are very lucky, we may have a situation where there is only *one* alternative.) Conventionally, we denote the null hypothesis by H_0, and its parameter value by θ_0. (Note that θ_0 no longer means the *true* parameter, since that's precisely the point at issue!)

There are three ways of thinking about the null.

11.2.1 The *Status Quo* Null

This sort of null is what we actually believe. We want to see whether we have to go through the agony of changing our minds, or whether we can continue in our dogmatic slumbers.

One obvious problem with *status quo* nulls is that sometimes we're not sure what we believe; we may not believe anything, really. A more subtle problem is that, even if we believe a certain hypothesis holds, for technical reasons of test-construction we may not want to make it the null.

11.2.2 The It-Would-Be-the-Worst-Mistake Null

Suppose we've got to choose between two courses of action as the result of our test — we do one thing if we reject the null, and another if we keep it. We assume that there's some cost if we choose the wrong course of action. Then we make the null the hypothesis whose action is most costly if chosen wrongly.

Let me unpack that a little. Think of medical testing for some rare disease: either the patient hasn't or he doesn't. If we think he does, we stick him full

of some chemical which will combat the disease if it's present, but makes him wretchedly sick if it's not. If we don't think he has the disease, we let him alone, and he gets sick if we're wrong. If false negatives — letting the disease go untreated — are worse than false positives — giving the medicine to a healthy person — then we say that the null is not having the disease.

The reason for this is that, as we'll see, we can control the chance of wrongly rejecting the null much more easily than we can control the chance of wrongly accepting the rival.

This sort of null is very important in signal detection theory, behavioral ecology (Shettleworth 1998), and so forth. The problem with using it to evaluate scientific hypothesis is that it's hard, often, to know which course of action would be worse for us, if we shouldn't have done it. (There is also the what-you-mean-we-white-man objection: worse for *who*?) You can, with some sophistry, bring *status quo* nulls under this heading, but that's not of much use to us.

11.2.3 The Random-Effects Null

Here we want to see whether our data could be due to chance, or more generally to some boring, uninteresting, stop-the-investigation-now mechanism. In other words, the idea is that what looks interesting about our data might just be a mistake. So we craft a null hypothesis which embodies this mistake, and if we can reject it then we think that the data were *not* due to chance, or more generally to boring mechanisms.[2]

This is my favorite kind of null, because we *can* apply it straightforwardly to scientific questions, and because I have some skill in making up null models. It's important to remember that in this case, each hypothesis test is a test against a *specific* kind of error. If we can be confident that we've avoided one sort of mistake, that doesn't mean that we've not made many others!

11.2.4 The Alternative Hypothesis

In formal hypothesis testing, we always have at least one rival to the null in mind. In the easiest cases, it's another distribution of the same form, but with a different parameter. Or it could be of a completely different form. If there is only one alternative acceptable to us, we write it H_1 and its parameter (if applicable) θ_1. This is the case of a **simple alternative**. If, on the other hand, there are several acceptable rivals — a **composite alternative** — we index them either numerically (if there's a finite number of them) or by their parameters (if not).

The null and the alternatives together make up the **class of admissible hypotheses**.

[2] *For specialists.* I'd be willing to argue that the maximum entropy principle (Jaynes 1983) is *really* a successful recipe for crafting null models. But developing that argument here would require, at the least, a chapter on information theory.

11.3 Formalism of Tests

11.3.1 The Test Statistic

Call the test statistic T.

Goodness-of-fit measures are common test statistics, but they're not the only ones. Generally speaking, we do chose test statistics for which increasingly large values are increasingly improbable, if the null is true, but even that isn't strictly necessary. All we really need to do is be able to calculate the sampling distribution of the test statistic under both the null and the alternative hypothesis.

11.3.2 The Regions

We divide values of T into two sets: the acceptance region $\mathcal{X}_0$, where we keep the null, and the rejection region $\mathcal{X}_1$, where we reject it in favor of the alternative. These regions need not be contiguous. (This is particularly true if the test statistic is actually multidimensional.) Some authors speak of the **critical region** as the boundary between the two, but they differ as to whether we should accept or reject if we hit right on the critical region, and the concept doesn't seem necessary.

The problem of designing a good test is to place the acceptance region so that we minimize our error probabilities.

11.3.3 The Kinds of Errors; Error Probabilities

There are two types of errors we can make in a test. The null hypothesis could be true, and we could still reject it (**Type I error**); or one of the alternative hypothesis could be true, and we could still accept the null (**Type II error**).

The error probabilities of a test are sometimes called its **operating characteristics**.

Significance Level or Size

The **significance level** of a test, conventionally α, is its probability of making a type I error, of falsely rejecting the null. It's the probability of getting a value of the test statistic inside the rejection region, calculated *as though the null were true*:

$$\alpha = \mathrm{P}_{H_0}(T \in \mathcal{X}_1) \ . \tag{11.3}$$

Suppose we use a test with some low value for α, e.g. 0.002, and we get a rejection. Then we know, regardless of what the data were, that it was rather unlikely we would have rejected the null if it were true; the test is good at avoiding false negatives. Note however that, if we had gotten results in the acceptance region instead, knowing the significance level of the test wouldn't have told us much, if anything, about how trustworthy that acceptance was.

Power

The **power** of a test, conventionally β, is the probability of not making a type II error, that is, the probability that the null will not be accepted when the alternative hypothesis is true:

$$\begin{aligned} \beta &\equiv \mathrm{P}_{H_1}(T \notin \mathcal{X}_0) && (11.4) \\ &= \mathrm{P}_{H_1}(T \in \mathcal{X}_1) \ . && (11.5) \end{aligned}$$

Power is in general a function of which alternative hypothesis we consider; tests which have a high power against one alternative may have a low power against another. If we are dealing with a composite alternative, we write β as a function of which hypothesis it is we're considering. In some cases, we can choose the acceptance region so as to maximize power against *all* the allowed alternatives; this gives us what are called **uniformly most-powerful tests**. They're good things when you can find them.

Power can be hard to calculate analytically, even when you know *exactly* the alternatives you are considering. But then we can always use Monte Carlo.

Severity

We do our experiment and we get a particular value for the test statistic, call it $\hat{t}$. We either accept the null or we reject it. Suppose we accept. Then we could ask, what is the probability that a particular alternative hypothesis would have given us a value of a test statistic which matches the null at least as well as our data did? Assume that, under the null, increasingly large values of T become increasingly improbable, so $\mathcal{X}_0$ is just all values of t below some threshold t_c. Then we compute, given that $\hat{t} \leq t_c$, for a specific H_i,

$$\mathrm{P}_{H_i}(T < \hat{t}) \ . \tag{11.6}$$

This is the probability of passing spuriously; one minus this is the **severity** of the test which H_0 passed against H_i. It is the probability that we would not have gotten such a good fit as we did, *if the null were false*. The severity against a bunch of H_i is the minimum of the individual severities. Similarly, for rejections, we ask how likely it is that we could get an even worse fit than the one which made us reject the null, even if it is true.

Note that the severity of accepting the null, vs. H_i, is *at least* the power of the test against H_i.

The severity (of a passing result) is not the probability that the null is true. It is the probability that, if the null is false, this test, fed this data, would have caught it. It tells us whether or not we have a high probability of having avoided a certain kind of mistake. I think severity is a very important concept, certainly when it comes to ruling out mistakes, but curiously unappreciated. Readers are referred to Mayo's book (1996), and her paper with Spanos (2000) for details.

The Trade-Offs

It's always possible to design very bad tests, which have both very high significance levels and low power; these are tests which can, so to speak, be improved in both directions at once. Unfortunately, it's generally not possible to design tests with both a significance level of 0 and a power of 1, so at some point we hit a limit where we need to begin making trade-offs: more power for higher significance levels, or vice versa. Here unfortunately we come to matters which can only be decided by experience and the particulars of the problems you face, i.e., taste and caprice.

On the other hand, the use of some significance levels is standard: social-science journals generally don't bother with anything which isn't significant at the five percent level, for instance. A standard significance level, plus a standard test, will fix your procedure.

11.3.4 Test for Whether Two Sample Means Are Equal

Let's put these ideas together to come up with a test, in a reasonably simple situation. We have two heaps of one hundred IID samples each, giving us two sample means, $\hat{\mu}_1 = 10.2$ and $\hat{\mu}_2 = 10.0$. Now, $\hat{\mu}_1 \neq \hat{\mu}_2$, but we know that two samples from the same population may not have the same mean simply by chance. So our random-errors null H_0 is the hypothesis that $\mu_1 = \mu_2 = \hat{\mu}$, where $\hat{\mu}$ is the sample mean from pooling both samples = 10.05. That is, we are looking at two samples from populations with the same mean, and we estimate that mean from the overall sample. The alternatives are simply summed up by $\mu_1 \neq \mu_2$; we'll index them by $\theta = \mu_1 - \mu_2$. Now, we know from the CLT that the sample means should be approximately Gaussian, and we know that the sum (or difference) of two independent Gaussians is another Gaussian. So if H_0 is true, $\hat{M}_1 - \hat{M}_2 \sim \mathcal{N}(0, 2\hat{\sigma}^2)$, where $\hat{\sigma}^2 = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}$ is the population variance as estimated from the sample. On the other hand, if H_θ is true, we expect $\hat{M}_1 - \hat{M}_2 \sim \mathcal{N}(\theta, 2\hat{\sigma}^2)$. (That is to say, our class of admissible hypotheses limits itself to those where the variances of the two populations are equal.) Say $\hat{\sigma}_1 = 0.097$ and $\hat{\sigma}_2 = 0.101$, so that $\hat{\sigma} = 0.099$. This suggests that a nice test statistic would be $Z = \frac{\hat{M}_1 - \hat{M}_2}{\sqrt{2}\hat{\sigma}}$, which $\sim \mathcal{N}(0,1)$ under H_0 and $\sim \mathcal{N}(\frac{\theta}{\sqrt{2}\hat{\sigma}}, 1)$ under H_θ.

Clearly, Z should have a small absolute value if H_0 is true, and its most likely value is 0. It can be shown that, for the alternatives we've chosen, the uniformly most powerful test has for its acceptance region a symmetric interval centered at zero, so we'll use that. A customary significance level is $\alpha = 0.05$, and this says that the acceptance region is $-1.96 \leq z \leq 1.96$. Our actual value of $z = \frac{10.2 - 10.0}{(\sqrt{2})(0.099)} = 1.43$, and so the null is saved.

How severe a test was this? Well, that depends on the alternative we have in mind. For H_θ, we need to compute $S(\theta) = 1 - \mathrm{P}_{H_\theta}(|Z| \leq 1.43)$, the probability of getting a fit to H_0 at least as good as what our data gave us. Now we saw that $Z \sim \mathcal{N}(\frac{\theta}{\sqrt{2}\hat{\sigma}}, 1) = \mathcal{N}(7.14\theta, 1)$, so $Z' = Z - 7.14\theta \sim \mathcal{N}(0,1)$. So

$S(\theta) = 1 - \mathrm{P}(-1.43 - 7.14\theta \leq Z' \leq 1.43 - 7.14\theta)$. There isn't a nice analytical expression for this, so I've tabulated it for a few values of θ.

θ	$S(\theta)$
± 0.05	0.179
± 0.10	0.255
± 0.20	0.514
± 0.30	0.761
± 1.00	≈ 1

Notice that, while we were able to reject $H_{0.05}$, the severity is low; indeed, if $H_{0.05}$ were true, the chance of getting data that fit the null at least as well as our data do is better than four-fifths.

Chapter 12

Funky Statistics

Or, more advanced stuff you might find useful or interesting.

12.1 Nonparametric Estimation and Fitting

What we've talked about assumes that you know the *kind* of distribution involved, and you're just not sure what the parameters are. But maybe you don't have such luck. There are actually techniques which don't require this. The Kolmogorov-Smirnov test we mentioned is a non-parametric test for whether or not two samples come from the same distribution. Mostly, though, non-parametric methods are useful in estimation and curve-fitting. Artifical neural networks are (in some of their guises) non-parametric curve-fitters — see Zapranis and Refenes (1999). So are the piecewise-polynomial functions called "splines," for which there is a very elegant theory (Wahba 1990). There are others, with odd names like "kernel density estimators" and "support vector machines" and so forth. Most of them assume some *sort* of functional form for the distribution you care about, but promise that you can approximate any distribution you like arbitrarily closely. Generally, there's a trade-off here between having simple non-parametric models and having accurate ones...

What computer programmers call "data-mining" (Weiss and Indurkhya 1998) is extremely relevant here, though not enough is being done at the interface.

12.2 Machine Learning

Lots of machine learning problems are formally problems of using algorithms to do estimation (possibly non-parametric). This is extremely fruitful in actually building up a theory of machine (and animal) learning, and in turn leads to new sorts of statistical problems. See Kearns and Vazirani (1994), Vapnik (2000) (though that's a very *Russian* mathematical tome), and Valiant (1994).

12.3 Causal Inference

Correlation is not causation. Not even statistical dependence is causation; at best both of these tell us that there's an association between two variables, but that could be cause one causes the other, or vice versa, or both are caused by other things and they have causes in common. In the last ten years or so, there's been a lot of work on the conditions under which we *can* legitimately infer causal models from statistical data, and what the inference procedures should be. See Pearl (2000) or Spirtes, Glymour and Scheines (2001) (a second edition is supposed to come out this summer), for regression-type problems, and Shalizi and Crutchfield (2001) for causal models of processes.

12.4 Ecological Inference

Can we recover any information about individual actions from aggregate data? More than might be thought. See King (1997).

12.5 Optimal Experimental Design

If we want to do estimation or testing, not every selection of independent variables is equally good. Depending on what you want to do, there is actually mathematical theory about which distributions are better than others, and what the best optimal ones are. This can be *extremely useful* if you are planning experiments or simulations which involve collecting large amounts of data. See Atkinson and Donev (1992).

Chapter 13

A Stochastic Process Is a Sequence of Random Variables

Or a random variable whose value is a sequence; the latter view is actually a little easier to handle mathematically. It's common to write the successive random variables $S_0, S_1, \ldots S_t, \ldots$ and so forth. Now, the space each of these random variables lives over is the same, and when we need to talk about that we'll talk about $\mathcal{S}$, and realizations will be s.

Distributions are now over sequences. Sometimes we imagine these to start at some point in time and then go on forever, and at others to go from minus infinity to plus infinity. Distributions over finite-length sequences, starting and stopping at a certain time, are marginal distributions from these. That is, if we want to know the distribution over sequences running from t_0 to t_1, we sum over the variables which represent what happened before t_0 and what will happen after t_1.

We could probably stand more convention about how to write finite sequences than we have; I'll use $\overrightarrow{S_t}^L$ for the sequence of length L which starts from, and includes, S_t. Similarly, $\overrightarrow{S_t}$ is the infinite series starting at t.

I've been talking, all this time, as though time were discrete. This is by far the easier case to handle. If time is continuous, then stochastic processes are distributions over *functions*, and that's mathematically very tricky to handle. I'll try to say a little bit about continuous-time stochastic processes at the end, but I'm afraid it'll only be a little.

13.1 Representing Stochastic Processes with Operators

Consider some particular value of $\vec{S}_t$, call it $\vec{s}_t = s_t s_{t+1} s_{t+2} \ldots$. What is $\vec{s}_{t+1}$? Well, obviously just $s_{t+1} s_{t+2} \ldots$. To move one time-step into the future, we just lop off the first item in the sequence. The operator $\mathbf{T}$ which does this is called the **shift** or **shift map**, and we say $\mathbf{T}\vec{s}_t = \vec{s}_{t+1}$. Its inverse, $\mathbf{T}^{-1}$, gives all the trajectories which could map to the trajectory we apply it to: $\mathbf{T}^{-1}\vec{s} = \left\{\vec{s}' | \mathbf{T}\vec{s}' = \vec{s}\right\}$. We say that $\mathbf{T}^{-1}\vec{s}$ consists of the **predecessors** of $\vec{s}$.

We can apply $\mathbf{T}$ and $\mathbf{T}^{-1}$ to a set of trajectories A; the result is the union of applying them to the individual trajectories, i.e., $\mathbf{T}A = \bigcup_{\vec{s} \in \mathrm{A}} \mathbf{T}\vec{s}$, and likewise for $\mathbf{T}^{-1}$. $\mathbf{T}$A is the **image** of A, and $\mathbf{T}^{-1}$A its **inverse image**.

A set of trajectories is **invariant** if $\mathbf{T}$A = A, if it is its own image. (Then it's its own inverse image too.)

Suppose we have a certain distribution μ_t over trajectories at time t. What will be the distribution at the next time step, i.e., what will be $\mu_{t+1} = \mathbf{T}\mu_t$? It is determined by the obvious equation,

$$\mathrm{P}(\vec{S}_{t+1} \in \mathrm{A}) \quad = \quad \mathrm{P}(\vec{S}_t \in \mathbf{T}^{-1}\mathrm{A}) \; . \tag{13.1}$$

This is a very crude introduction to the important topic of symbolic dynamics; see Badii and Politi (1997), Beck and Schlögl (1993) or Kitchens (1998). (The last is *very* mathematical.)

13.2 Important Properties of Stochastic Processes

13.2.1 Stationarity

Stationarity is simply the property of time-translation invariance for the distribution. More exactly, if we may abuse notation by writing $\mathrm{P}(\vec{S}_t^L)$ for the distribution of $\vec{S}_t^L$, then the process is stationary if and only if $\mathrm{P}(\vec{S}_t^L) = \mathrm{P}(\vec{S}_{t+\tau}^L)$, for any integer τ, for all L.

Observe that if S_t is stationary, then if $R_t = f(S_t)$, for some deterministic function f, the sequence R_t is also a stationary process.

In terms of the shift operator of sec. 13.1, if the process is stationary, then the distribution μ over trajectories is unaffected by the shift: $\mathbf{T}\mu = \mu$. The converse is also true.

13.2.2 Ergodicity

A process S_t is **ergodic** if there exists a random variable Y such that $\mathbf{E}S_1 = \mathbf{E}Y$ and, with probability one,

$$\frac{1}{T}\sum_{i=1}^{T} S_i \quad \xrightarrow[T\to\infty]{} \quad Y \ . \tag{13.2}$$

The left hand side of Eq. 13.2 is called the **time average** of S_t.

Theorem. Any stationary process such that $\mathbf{E}|S_1| < \infty$ is ergodic.

If the distribution over trajectories μ (see sec. 13.1) is such that, for every invariant set of trajectories A, either $\mathrm{P}_\mu(\mathrm{A}) = 1$ or $\mathrm{P}_\mu(\mathrm{A}) = 0$, then $Y = \mathbf{E}S_1$ almost surely.

The above is for discrete time; in continuous time, replace the sum with the left hand side of Eq. 13.2 with $\frac{1}{T}\int_0^T S(t)dt$.

13.2.3 Mixing

The mixing property is essentially that of forgetting what happened in the past. It gets its name from the idea that, if you take a well-defined chunk of the state space and let it evolve, it will get mixed with the images of any other nice chunk.

More technically, a quantity conventionally called α is defined as the maximum deviation from independence between any two events at two different times:

$$\alpha(t,t') \quad \equiv \quad \max_{\mathrm{B,C}} |\mathrm{P}(S_t \in \mathrm{B}, S_{t'} \in \mathrm{C}) - \mathrm{P}(S_t \in \mathrm{B})\mathrm{P}(S_{t'} \in \mathrm{C})| \tag{13.3}$$

If α decays exponentially as $t-t'$ grows, then the process is **strongly α-mixing**, and various powerful statistical techniques can be applied to it (Bosq 1998). The strong-mixing property is also thought to be important in the foundations of statistical mechanics.

Chapter 14

Markov Processes

In deterministic dynamics, future states are uniquely fixed by the present state alone; how that state was reached is irrelevant. If our model has this property, that is a good (but hardly infallible) sign that we've found the right states. It would be nice to extend this notion to stochastic processes, and that is done through the idea of a Markov process.

A discrete-valued, discrete-time stochastic process has the **Markov property** when

$$\begin{aligned} &\mathrm{P}(S_{n+1} = s_{n+1} | S_n = s_n) = \\ &\quad \mathrm{P}(S_{n+1} = s_{n+1} | S_n = s_n, S_{n-1} = s_{n-1}, \ldots S_1 = s_n) \end{aligned} \tag{14.1}$$

for all s_i and for all n. That is, the probability distribution for the next state depends solely on the current state. We then say that S_n is a **Markov process**. The analogous condition for continuous time is

$$\begin{aligned} &\mathrm{P}(S(t_{n+1}) = s_{n+1} | S(t_n) = s_n) = \\ &\quad \mathrm{P}(S(t_{n+1}) = s_{n+1} | S(t_n) = s_n, S(t_{n-1}) = s_{n-1}, \ldots S(t_1) = s_1) \end{aligned} \tag{14.2}$$

for all values s_i and any increasing sequence of times t_i.

Finally, a real-valued, continuous-time stochastic process is Markovian when

$$\begin{aligned} &\mathrm{P}(X(t_{n+1}) \leq x_{n+1} | X(t_n) = x_n) = \\ &\quad \mathrm{P}(X(t_{n+1}) \leq x_{n+1} | X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \ldots X(t_1) = x_1) \end{aligned} \tag{14.3}$$

for any x_i and any increasing sequence of times t_i. This only requires that $X(t)$ have a CDF at every point in time, not necessarily that it have a pdf.

14.1 Markov Chains and Matrices

A discrete-valued Markov process, whether in discrete or continuous time, is called a **Markov chain**. Let's think about the discrete-time case first.

We know, from the definition, that the important quantities are the probabilities $\mathrm{P}(S_{n+1} = j|S_n = i)$. If these probabilities are independent of n, then we say that the chain is **homogeneous**. In that case, we represent the probabilities by the elements of the **transition matrix** $\mathbf{T}$, where $T_{ij} = \mathrm{P}(S_{n+1} = j|S_n = i)$, the probability of going from state i to state j. It is often easiest to figure out what the process will do by examining the properties of its transition matrix.

If we write the distribution over states at one time as a column vector μ, then the distribution over states at the next time-step is just $\mathbf{T}\mu$. If we know that we are in state i now, and wish to know the probability of being in state j after n time-steps, that is simply the ij^{th} element of $\mathbf{T}^n$, which I will write as T^n_{ij}.

The **images** of a state i are the points it could go to next, i.e., the j such that $T_{ij} > 0$. The **inverse images** of i are the points which could have led to i, the j such that $T_{ji} > 0$. The (inverse) image of a set of points is the union of their (inverse) images. A set is **invariant** if it is equal to its own image.

14.2 Some Classifications of States, Distributions and Chains

A state i is **recurrent** or **persistent** if there is some $n \geq 1$ such that $T^n_{ii} = 1$; that is, after some amount of time, a process which starts in i is bound to return to i. If a state is not recurrent it is **transient**. A state is **absorbent** or an **absorbing state** if $T_ij = 0$ for all $j \neq i$.

The **mean recurrence time** of state i is the expectation-value of the time required to go from i back to itself. A persistent state with infinite mean recurrence time is **null**, otherwise it is **non-null**.

The **period** k of state i is the greatest common denominator of the n such that $T^n_{ii} > 0$. If $k = 1$ the state is **aperiodic**, otherwise it is **periodic**. A persistent, non-null, aperiodic state is **ergodic**.

In general, all these terms apply to sets of states if they apply to each member of the set.

State i **communicates** with state j if, for some n, $T^n_{ij} > 0$; that is, it is possible to go from i to j. If j also communicates with i, they **intercommunicate**.

A set of states C is **irreducible** if, for every pair of states in C intercommunicates. The whole state space may be irreducible.

A set of states C is **closed** if no state in C communicates with any state outside C. This is equivalent to having $T_{ij} = 0$ for all $i \in C$ and any $j \notin C$. (An absorbing state is a closed set with only one member.)

A distribution over the state space is **invariant** if it is mapped to itself, i.e. if $\mathbf{T}\mu = \mu$. Thus the invariant distributions are the eigenvector of the matrix $\mathbf{T}$ with eigenvalue 1, and they can be found by solving for those eigenvectors. A set of states is invariant if it is closed and irreducible. An invariant distribution is **ergodic** it it gives every invariant set either probability 0 or probability 1.

Theorem. An irreducible chain has a unique invariant distribution iff all its states are non-null persistent.

If a chain is irreducible and "ergodic", i.e., every state is persistent, non-null and aperiodic, then the ergodic theorem applies.

Theorem (Ergodic Theorem for Markov Chains). If an irreducible aperiodic chain has a unique invariant distribution μ_j, then $T^n_{ij} \to \mu_j$ as $n \to \infty$. Furthermore, $\mathrm{P}(X_n = j) \to \mu_j$.

14.3 Higher-Order Markov Chains

Suppose that

$$\begin{aligned} \mathrm{P}(S_{i+2}, S_{i+1}, S_i) &\neq \mathrm{P}(S_{i+2}|S_{i+1})\mathrm{P}(S_{i+1}|S_i)\mathrm{P}(S_i) \text{ but} && (14.4) \\ \mathrm{P}(S_{i+3}, S_{i+2}, S_{i+1}, S_i) &= \mathrm{P}(S_{i+3}|S_{i+2}, S_{i+1}) \\ &\quad \times \mathrm{P}(S_{i+2}|S_{i+1}, S_i)\mathrm{P}(S_{i+1}, S_i) && (14.5) \end{aligned}$$

The process isn't a Markov chain, but the future depends only on the past *two* values; it would be a chain if we looked at *pairs* of values. (*Exercise*: show that.) So we say that this is a **second-order** Markov process. If we need to condition on the n previous variables, then it's an n^{th}**-order** process. If you can't have a Markov chain, then having a higher-order Markov process is often the next best thing.

14.4 Hidden Markov Models

If we apply a function to the state of a Markov chain, or the transition between states, we get a new stochastic process (just like applying a function to an ordinary random variable gives us a new random variable). In general, this new process will *not* have the Markov property. This suggests that if we have data which isn't Markovian, we might still be able to describe it as a function of a Markov chain, and then we'll be happier.[1] Such a beast is called a **hidden Markov model** or **HMM**, and there are well-established algorithms for inferring them from data. Each transition between the hidden states now has a value of our observed data tacked on: when the process makes that transition, we see that value in our data-stream. (Formally, we now have one transition matrix for every value of the data.)

There is a friendly introduction to inferring HMMs in Charniak (1993); Elliott, Aggoun and Moore (1995) is considerably more advanced.

[1] One of the theorems in Shalizi and Crutchfield (2001) shows that every stationary discrete-valued, discrete-time process can be represented in this way.

Chapter 15

Examples of Markov Processes

15.1 Bernoulli Trials

This is dead-stupid, but it fits all the definitions...

15.2 Biased Drift on a Ring

Consider a Markov chain with five states and the following transition matrix:

$$\mathbf{T} = \begin{bmatrix} .2 & .7 & 0 & 0 & 0.1 \\ .1 & .2 & .7 & 0 & 0 \\ 0 & .1 & .2 & .7 & 0 \\ 0 & 0 & .1 & .2 & .7 \\ .7 & 0 & 0 & .1 & .2 \end{bmatrix}$$

This represents motion along a ring, with a general trend in one direction, but sometimes staying put or drifting the other way. The whole space is irreducible, every state is non-null, persistent and aperiodic, and the invariant distribution is the uniform distribution.

15.3 The Random Walk

A.k.a. the drunkard's walk. The probability space is the number-line. At each time-step, we either add one to our current location, or subtract one; this is completely independent of everything. This is a sort of Bernoulli variable, only taking values from $\{-1, +1\}$ rather than $\{0, 1\}$. Start at the origin at time 0. Then the position at time N depends only on how many plus moves (m) there have been, since the others must've been minus moves. If we have had m plus

moves, then our position $= 0 + m - (N - m) = 2m - N$. Or, if $r = 2m - N$, $(r + N)/2 = m$. The probability of being at position r after N moves is thus given by a binomial probability, that of having $m = (r + N)/2$ successes out of N chances.

$$P(X(N) = r) = \left(\begin{array}{c} N \\ (r+N)/2 \end{array} \right) (0.5)^{(r+N)/2} 0.5^{N-(r+N)/2} \quad (15.1)$$

$$= \frac{1}{2^N} \left(\begin{array}{c} N \\ (r+N)/2 \end{array} \right) \quad (15.2)$$

Exercise: Convince yourself that the random walk has the Markov property.

As $N \to \infty$, the distribution of the position of the random walker approaches a normal distribution, $\mathcal{N}(0, N)$. The mean distance from the origin is always 0, because we're equally likely to go be on either side of it. The expectation of the square of distance from the origin is the variance plus the square of the mean, or N. So in a way we drift away from the origin by a distance which grows as the square-root of the time we allow the process to run. This is despite the fact that, at each step, we are just as likely to move towards the origin as away from it.

Naturally, this generalizes to two or more dimensions: instead of just having the moves $\{-1, +1\}$, we let the random walker move along each of the axes with equal probability. Then the exact distribution for the coordinates is a product of binomial distributions (each coordinate is an independent random variable), and the limit is again Gaussian, with a square-root drift from the origin. In one dimension, the random walk can be expected to return to the origin infinitely often; this is not true in higher dimensions.

To a good approximation, physical processes of diffusion also have this square-root dependence on time, and random walks make good models for them; this is part of why they have been studied to death. (Hughes's *Random Walks and Random Environments* (1995), for instance, fills two volumes of about 800 pages each, and is far from complete.)

Chapter 16

Continuous-Time Stochastic Processes

16.1 The Poisson Process

"Hits" arrive randomly and accumulate. $N(t)$ is the number of hits which have arrived up to time t; $N(0) = 0$. The time between successive hits is exponentially distributed with parameter λ, and the inter-arrival times are independent of each other. Then $N(t)$ is Poisson distributed with parameter λt, that is, $\mathrm{P}(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$.

The Poisson process is a Markov chain in continuous time, but with discrete states. Such beasts are sometimes called **jump processes**. Among the jump processes, the Poisson process is one of those which is said to be **cadlag**, an acronym for the French phrase ≪continue à droite, limite à gauche≫ — "continuous to the right, limited to the left". That is, we say that the function is continuous up to the jump point as we approach it from the right, and that it has a limit as we approach the jump point from the left. (Cadlag processes are also called **Skorokhod maps** and **R-processes**.)

Because $N(t)$ is a process in continuous time, we cannot just represent it with a simple transition matrix as we could discrete-time Markov chains. Of course we can define the transition probabilities $p_{ij}(s,t) = \mathrm{P}(N(t) = j | N(s) = i)$, and these are (by the Markov property) all we need. If $p_{ij}(s, s+\tau) = p_{ij}(0, \tau)$, then the process is **homogeneous** (just as in the discrete case), and we can represent that set of transition probabilities by a matrix $\mathbf{T}_\tau$. It is very often (but not universally) the case that there is a **generator**, a matrix $\mathbf{G}$ such that $\mathbf{T}_\tau = e^{\mathbf{G}t}$. In particular, this is the case for the Poisson process, where the entries of G are given by

$$G_{ij} = \begin{cases} -\lambda & \text{if } j = i \\ \lambda & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases} .$$

16.1.1 Uses

The Poisson process is a good model for random "arrivals" — the number of nuclei of a radioactive substance which have decayed in a certain amount of time, for instance. It makes a good null model for neural spike trains (Rieke, Warland, de Ruyter van Steveninck and Bialek 1997), and, generalized to allow differences in *where* "hits" occur, it turns out to be very good at describing how things (raindrops, hail, WWII-vintage bombs) fall from the sky.

In a sense which can be made precise through maximum-entropy arguments, the Poisson distribution is the "most random" way of arranging a specified mean number of marks in a fixed interval.

16.2 Brownian Motion, or the Wiener Process

This is the continuous-time generalization of the random walk. The paths it describes are fully continuous, and it was the first completely continuous stochastic process with a mathematically decent theory behind it, which Wiener worked out in the 1920s. In the '40s he adapted the same math to problems of control (feedback mechanisms), prediction (automatic anti-aircraft guns) and communication (radio and radar), and came up with cybernetics, whence (by a long route) us here today. The uses of pure mathematics, when you throw a lot of military money at it, are wonderful to behold...[1]

$W(t_2) - W(t_1)$ is the increment between t_2 and t_1. The condition for the Wiener process is that

1. *Stationarity.* For any two times $t_1 < t_2$, the increment $W(t_2) - W(t_1)$ depends only on $t_2 - t_1$.

2. *Independence.* For any three times $t_1 < t_2 < t_3$, $W(t_3) - W(t_2)$ and $W(t_2) - W(t_1)$ are independent.

3. *Normality.* $W(t_2) - W(t_1)$ is normally distributed.

Using these three, you can show that $W(t_2) - W(t_1)$ has the density function $\mathcal{N}(0, \sigma^2\tau)$, where $\tau = t_2 - t_1$, and σ is a parameter we're free to vary.

The curves described by the Wiener process are continuous, and in fact Wiener's prescription gives us a way of assigning probabilities on the whole space of continuous curves. (There is a very nice construction of this probability measure in the opening chapters of Wiener's book on *Nonlinear Problems in Random Theory* (1958), which otherwise is the kind of thing that makes my head want to explode.) Unfortunately, the curves are almost never differentiable, i.e., they almost never have a well-defined derivative with respect to time.

Suppose we start the Wiener process at the origin: $W(0) = 0$. Then we can ask where it is, on average, at later times. Well, the average position at t is just $\mathbf{E}W(t) = \mathbf{E}(W(t) - W(0)) = 0$, since increments are normally-distributed

[1] Wiener (1961) describes the earlier part of this history.

about zero. But if we look at the square of the position, $\mathbf{E}((W(t) - W(0))^2)$, that is the variance plus the square of the mean, or $\sigma^2 t$. (Remember we got a result like this from the random walk.) So if we try to compute an average speed on that basis, we get $\bar{v} = \frac{\sqrt{\sigma^2 t}}{t} = \frac{\sigma}{\sqrt{t}}$. And if we take the limit $t \to 0$, this blows up — which means that the curve doesn't have a well-defined derivative.

The physical picture here is that, as we look over briefer and briefer times, we see incredibly intense activity (high $\bar{v}$) which, however, has absolutely no direction to it; most of that motion cancels out. Enough is uncanceled, over finite times, that we tend to drift away from the origin, but more slowly than we would if we had any non-zero, constant velocity.

If the Wiener process *had* a derivative, it would be white noise — the noise with a flat power-spectrum, whose values at any two times are independent, identically-distributed Gaussians. This can actually be made quite precise, but involves horrendous excursions through analysis and measure-theory to prove; Gardiner's book (1990) on stochastic methods and Keizer's book on molecular fluctuations (1987, ch. 1) are about the most hand-holding, cook-booky ways of approaching it. (Keizer perversely writes his conditional probabilities backwards, but it's an excellent book otherwise, and has a nice treatment of more realistic models of Brownian motion than the Wiener process, i.e., ones where the velocity is well-defined.)

Appendix A

Notes for Further Reading

A.1 Probability

A good textbook and reference (i.e., the one I usually look things up in first) is Grimmett and Stirzaker (1992). Stirzaker's solo effort (Stirzaker 1999) is intended for those with less math, or time, or interest. Feller (1957) is, deservedly, a classic. Billingsley (1995) is for mathematicians or masochists, but gives rigorous proofs of everything, which from an applied point of view often amounts to telling people that perfectly reasonable things will sometimes not work. There is a good discussion of how to define "random" algorithmically in Cover and Thomas (1991), among other places. Bennett (1998) is a popular book, roughly at the level of *Scientific American*, and nicely done.

Gigerenzer, Swijtink, Porter, Daston, Beatty and Krüger (1989), Hacking (1975, 1990), Porter (1986) and Stigler (1986) are all fine histories of the development of probabilistic methods and of statistics. Hacking in particular is an excellent writer.

For the debates over what probabilities should *mean*, see below under statistics.

A.2 Statistics

The classic work on mathematical statistics, which includes a superb discussion of measure and probability theory, is Cramér (1945). There are any number of competent and recent textbooks, of which I might name Lupton (1993). Reid (1982) is an excellent biography of one of the founders of modern statistics. Gonick's book (Gonick and Smith 1993) is, like all his works, excellent as far as it goes, but leaves out a lot.

The books by Tufte (1983, 1990, 1997) are masterpieces which cannot be too strongly recommended.

Lehmann (1997) and Lehmann and Casella (1998) are the Law and the Prophets for estimation and hypothesis-testing. Weisberg (1985) is a useful

introduction to linear regression. An interesting perspective on statistical inference is given by Kullback (1968).

Differences over whether probabilities should represent frequencies of occurence, degrees of belief, or something else altogether make real differences in the kind of statistics you do. People who like the degrees-of-belief idea will find the kind of estimation and hypothesis-testing I've presented completely pointless, and vice-versa. A classic presentation of the degrees-of-belief school, in its most modern incarnation called Bayesianism, is Savage (1954). Bernardo and Smith (1994) is an encyclopedia. The anti-Bayesian arguments which convinced me — and a very good book in general — are those of Mayo (1996), which is also highly recommended if you have any interest in the philosophy of science. Rissanen (1989) makes an interesting (if ultimately unconvincing) case for making probabilities refer not to frequencies or beliefs but to coding schemes; the book is hard to follow unless you already know information theory.

The third traditional branch of statistical inference, besides estimation and testing, is decision theory. Luce and Raiffa (1957) is a good compendium of ideas thereon, though with a Bayesian bias (i.e., they tend to assume that it makes sense to say things like "I think the probability of Chelsea Clinton being elected President in 2044 is 0.01"). It's extremely well-established experimentally that human judgment and decision-making is not even close to what Bayesian theory would consider rational (Camerer 1995). Whether you think this is bad news for human beings (Kahneman, Slovic and Tversky 1982) or bad for Bayesian theory (Simon 1996; Cosmides and Tooby 1996; Gigerenzer, Todd and the ABC Research Group 1999) is to some degree a matter of taste.

I've already named relevant books about extensions to standard statistical inference.

A.3 Stochastic Processes

First of all, see Grimmett and Stirzaker, already mentioned; also, again, Feller. Hoel, Port and Stone (1987) is a decent modern textbook, which assumes a background probability course. Doob (1953) was the first modern book on stochastic processes, and still a standard reference for probabilists. Gardiner (1990) is a useful compendium of techniques and processes.

The literature on Markov processes is huge. All the textbooks I've just named devote considerable space to them. Kemeny and Snell (1976) and Kemeny, Snell and Knapp (1976) are encyclopedic.

Information theory is an important adjunct of stochastic processes, which will be covered at some depth during the course. Cover and Thomas (1991) is so much better and more complete than everything else on the subject that there's little point in reading any other book, unless you're going to make this your specialty. (In that case, however, see Kullback and Rissanen.)

Statistical inference for stochastic processes is a more scattered and tricky topic than might be expected. Bosq (1998) deals with powerful non-parametric techniques, but assumes considerable knowledge of statistics. Many books on

time-series effectively deal with aspects of the problem: see Kantz and Schreiber (1997), or, old but still valid and often useful, Grenander and Rosenblatt (1984), which is explicitly statistical. Billingsley (1961) collects useful results for Markov chains and jump-processes. Elliott, Aggoun and Moore (1995) is, like Billingsley, an advanced book, with many useful results and algorithms. Much of the machine-learning literature effectively talks about fitting stochastic processes from data, but too much of it ignores statistical issues of efficiency, bias, convergence, etc. (References which offend in this respect are available upon request.)

Appendix B

What's Wrong with These Notes

- There are no pictures.
- There are no real problems.
- There is no index or glossary.
- There isn't enough about how to actually *do* probability theory; the reader would have no idea about how to even sketch a proof of, say, the law of large numbers.
- There should be more on the binomial distribution and its manipulation, and at least a mention of the multinomial.
- The treatment of multi-dimensional continuous variables is very sketchy.
- Covariance should probably be moved to the section on moments.
- The chapter on data-handling needs to say more on display. Is there a graphics package which implements Tuftean principles?
- The chapters on data-handling and sampling may not be in the right order.
- The chapter on estimation should have some worked examples.
- I don't explain how to do linear regression.
- The chapter on hypothesis-testing doesn't say how to handle composite nulls; or say enough about calculating severity and its relations to size and power.
- It doesn't distinguish between one-sided and two-sided tests.
- In the Gaussian worked example, I don't show that the symmetric acceptance region delivers the uniformly most powerful test.

- There is no example using the χ^2 test.
- Describe the Kolmogorov-Smirnov test.
- I don't mention and refute any of the common fallacies about tests of hypotheses.
- There is nothing about decision theory; a chapter in Part II sounds about right.
- There is nothing about autocorrelation or power spectra for stochastic processes.
- There is nothing about information theory; a chapter in Part III sounds about right.
- There needs to be more on continuous-time stochastic processes. Define diffusion processes. Give a workable definition of white noise. Explain the Fokker-Planck equation.
- There is nothing on inference for stochastic processes. Stealing from Billingsley (1961), at the very least, suggests itself.

Bibliography

Atkinson, A. C. and A. N. Donev (1992). *Optimum Experimental Designs*. Oxford: Clarendon Press.

Badii, Remo and Antonio Politi (1997). *Complexity: Hierarchical Structures and Scaling in Physics*. Cambridge: Cambridge University Press.

Beck, Christian and Friedrich Schlögl (1993). *Thermodynamics of Chaotic Systems: An Introduction*. Cambridge, England: Cambr idge University Press.

Bennett, Deborah J. (1998). *Randomness*. Cambridge, Massachusetts: Harvard University Press.

Bernardo, Jose M. and Adrian F. M. Smith (1994). *Bayesian Theory*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Billingsley, Patrick (1961). *Statistical Inference for Markov Processes*, vol. 2 of *Statistical Research Monographs*. Chicago: University of Chicago Press.

— (1995). *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley, 3rd edn.

Bosq, Denis (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Berlin: Springer-Verlag, 2nd edn.

Camerer, Colin (1995). "Individual Decision Making." In *The Handbook of Experimental Economics* (John H. Kagel and Alvin E. Roth, eds.), pp. 587–703. Princeton, New Jersey: Princeton University Press.

Charniak, Eugene (1993). *Statistical Language Learning*. Cambridge, Massachusetts: MIT Press.

Cohn, Paul M. (1981). *Universal Algebra*, vol. 6 of *Mathematics and Its Applications*. Dordrecht, Holland: D. Reidel, 2nd edn.

Cosmides, Leda and John Tooby (1996). "Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgement Under Uncertainity." *Cognition*, **58**: 1–73.

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. New York: Wiley.

Cramér, Harald (1945). *Mathematical Methods of Statistics*. Uppsala: Almqvist and Wiksells. Republished by Princeton University Press, 1946, as vol. 9 in the Princeton Mathematics Series and as a paperback in the Princeton Landmarks in Mathematics and Physics series, 1999.

den Hollander, Frank (2000). *Large Deviations*, vol. 14 of *Fields Institute Monographs*. Providence, Rhode Island: American Mathematical Society.

Doob, Joseph L. (1953). *Stochastic Processes*. Wiley Publications in Statistics. New York: Wiley.

Elliott, Robert J., Lakhdar Aggoun and John B. Moore (1995). *Hidden Markov Models: Estimation and Control*, vol. 29 of *Applications of Mathematics: Stochastic Modelling and Applied Probability*. New York: Springer-Verlag.

Feller, William (1957). *An Introduction to Probability Theory and Its Applications*, vol. I. New York: Wiley, 2nd edn.

Gardiner, C. W. (1990). *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Berlin: Springer-Verlag, 2nd edn.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty and Lorenz Krüger (1989). *The Empire of Chance: How Probability Changed Science and Everyday Life*, vol. 12 of *Ideas in Context*. Cambridge, England: Cambridge University Press.

Gigerenzer, Gerd, Peter M. Todd and the ABC Research Group (1999). *Simple Heuristics That Make Us Smart*. Oxford: Oxford University Press.

Gillespie, John H. (1998). *Population Genetics: A Concise Guide*. Baltimore: Johns Hopkins University Press.

Gonick, Larry and Wollcoot Smith (1993). *The Cartoon Guide to Statistics*. New York: Harper Perennial.

Grenander, Ulf and Murray Rosenblatt (1984). *Statistical Analysis of Stationary Time Series*. New York: Chelsea Publishing, 2nd edn.

Grimmett, G. R. and D. R. Stirzaker (1992). *Probability and Random Processes*. Oxford: Oxford University Press, 2nd edn.

Hacking, Ian (1975). *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge, England: Cambridge University Press.

— (1990). *The Taming of Chance*, vol. 17 of *Ideas in Context*. Cambridge, England: Cambridge University Press.

Hoel, Paul G., Sidney C. Port and Charles J. Stone (1987). *Introduction to Stochastic Processes*. Boston: Houghton Mifflin.

Hughes, Barry D. (1995). *Random Walks and Random Environments*. Oxford: Clarendon Press.

Jaynes, E. T. (1983). *Essays on Probability, Statistics, and Statistical Physics*. London: Reidel.

Kahneman, Daniel, Paul Slovic and Amos Tversky (eds.) (1982). *Judgment Under Uncertainty*, Cambridge, England. Cambridge University Press.

Kantz, Holger and Thomas Schreiber (1997). *Nonlinear Time Series Analysis*. Cambridge, England: Cambridge University Press.

Kearns, Michael J. and Umesh V. Vazirani (1994). *An Introduction to Computational Learning Theory*. Cambridge, Massachusetts: MIT Press.

Keizer, Joel (1987). *Statistical Thermodynamics of Nonequilibrium Processes*. New York: Springer-Verlag.

Kemeny, John G. and J. Laurie Snell (1976). *Finite Markov Chains*. New York: Springer-Verlag.

Kemeny, John G., J. Laurie Snell and Anthony W. Knapp (1976). *Denumerable Markov Chains*. New York: Springer-Verlag, 2nd edn.

King, Gary (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, New Jersey: Princeton University Press.

Kitchens, Bruce P. (1998). *Symbolic Dynamics: One-sided, Two-sided and Countable State Markov Shifts*. Berlin: Springer-Verlag.

Kullback, Solomon (1968). *Information Theory and Statistics*. New York: Dover Books, 2nd edn.

Lehmann, E. L. (1997). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Berlin: Springer-Verlag, 2nd edn.

Lehmann, E. L. and George Casella (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Berlin: Springer-Verlag, 2nd edn.

Luce, R. Duncan and Howard Raiffa (1957). *Games and Decisions: Introduction and Critical Survey*. New York: Wiley.

Lupton, Robert (1993). *Statistics in Theory and Practice*. Princeton, New Jersey: Princeton University Press.

MacKeown, P. Kevin (1997). *Stochastic Simulation in Physics*. Singapore: Springer-Verlag.

Manoukian, Edward B. (1986). *Modern Concepts and Theorems of Mathematical Statistics*. Springer Series in Statistics. Berlin: Springer-Verlag.

Mayo, Deborah G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G. and Aris Spanos (2000). "A Post-Data Interpretation of Neyman-Pearson Methods Based on a Conception of Severe Testing." Preprint, Virginia Tech.

Newman, Mark E. J. and G. T. Barkema (1999). *Monte Carlo Methods in Statistical Physics*. Oxford: Clarendon Press.

Patel, Jagdish K., C. H. Kapadia and D. B. Owen (1976). *Handbook of Statistical Distributions*. New York: Marcel Dekker.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press.

Porter, Theodore M. (1986). *The Rise of Statistical Thinking, 1820–1900*. Princeton, New Jersey: Princeton University Press.

Press, William H., Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery (1992a). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, England: Cambridge University Press, 2nd edn. URL `http://lib-www.lanl.gov/numerical/`.

— (1992b). *Numerical Recipes in Fortran: The Art of Scientific Computing*. Cambridge, England: Cambridge University Press, 2nd edn. URL `http://lib-www.lanl.gov/numerical/`.

Pronzato, Luc, Henry P. Wynn and Anatoly A. Zhigljavsky (1999). *Dynamical Search: Applications of Dynamical Systems in Search and Optimization*. Interdisciplinary Statistics. Boca Raton, Florida: CRC Press.

Reid, Constance (1982). *Neyman from Life*. New York: Springer-Verlag.

Rieke, Fred, David Warland, Rob de Ruyter van Steveninck and William Bialek (1997). *Spikes: Exploring the Neural Code*. Cambridge, Massachusetts: MIT Press.

Rissanen, Jorma (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

Savage, Leonard J. (1954). *The Foundations of Statistics*. Wiley Pulibcations in Statistics. New York: Wiley.

Shalizi, Cosma Rohilla and James P. Crutchfield (2001). "Computational Mechanics: Pattern and Prediction, Structure and Simplicity." *Journal of Statistical Physics*, **104**: 817–879. URL `http://arxiv.org/abs/cond-mat/9907176`.

Shettleworth, Sara J. (1998). *Cognition, Evolution and Behavior*. Oxford: Oxford University Press.

Simon, Herbert A. (1996). *The Sciences of the Artificial*. Cambridge, Massachusetts: MIT Press, 3rd edn. First edition 1969.

Spirtes, Peter, Clark Glymour and Richard Scheines (2001). *Causation, Prediction, and Search*. Cambridge, Massachusetts: MIT Press, 2nd edn.

Stigler, Stephen M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, Massachusetts: Harvard University Press.

Stirzaker, David R. (1999). *Probability and Random Variables: A Beginner's Guide*. Cambridge, England: Cambridge University Press.

Tufte, Edward R. (1983). *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.

— (1990). *Envisioning Information*. Cheshire, Connecticut: Graphics Press.

— (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, Connecticut: Graphics Press.

Valiant, Leslie G. (1994). *Circuits of the Mind*. Oxford: Oxford University Press.

van de Geer, Sara (2000). *Empirical Processes in M-Estimation*. Cambridge, England: Cambridge University Press.

Vapnik, Vladimir N. (2000). *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 2nd edn.

Wahba, Grace (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Weisberg, Sanford (1985). *Applied Linear Regression*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: Wiley, 2nd edn.

Weiss, Sholom M. and Nitin Indurkhya (1998). *Predictive Data Mining: A Practical Guide*. San Francisco: Morgan Kaufmann.

Wiener, Norbert (1958). *Nonlinear Problems in Random Theory*. Cambridge, Massachusetts: The Technology Press of the Massachusetts Institute of Technology.

— (1961). *Cybernetics: Or, Control and Communication in the Animal and the Machine*. Cambridge, Massachusetts: MIT Press, 2nd edn. First edition New York: Wiley, 1948.

Zapranis, Achilleas and Apostolos-Paul Refenes (1999). *Principles of Neural Model Identification, Selection and Adequacy: With Applications to Financial Econometrics*. London: Springer-Verlag.