G. A. REYNOLDS

THINK STATS

A CONCEPTUAL INTRODUCTION TO STATISTICS

FOR THE SKEPTICAL, THE PESSIMISTIC, AND THE MILDLY DISTURBED

Contents

Introduction

What is statistics[1]?

[1] test note

> Modern statistics provides a quantitative technology for empirical science; it is a logic and methodology for the measurement of uncertainty and for an examination of the consequences of that uncertainty in the planning and interpretation of experimentation and observation.
>
> Stigler [1986]

> If all sciences require measurement–and statistics is the logic of measurement–it follows that the history of statistics can encompass the history of all of science.
>
> Stigler [1986]

Part I

Mathematic & Logic

Sets 'n Stuff

We use the Z Specification notation [2].

[2] ISO/IEC 13568:2002. *Information technology – Z formal specification notation – Syntax, type system and semantics.* ISO, Geneva, 2002. URL http://www.iso.org/iso/catalogue_detail.htm?csnumber=21573

Sets Membership, subset; family of sets

Relations

Functions

Sequences

Multisets

Sets

Notation: extension v. comprehension

Relations

Functions

Definition

Function Informally, a function is a set of ordered pairs. The Z specification says "A function is a particular form of relation, where each domain element has only one corresponding range element."[3]

[3] ISO/IEC 13568:2002. *Information technology – Z formal specification notation – Syntax, type system and semantics.* ISO, Geneva, 2002. URL http://www.iso.org/iso/catalogue_detail.htm?csnumber=21573

Function Extension Since a function is a kind of set, it can be defined by explicitly listing its extension–all of its elements. For example, the function f that maps each integer between 1 and 3 to itself can be expressed by writing out the complete list of its elements: $f = \{(1,1),(2,2),(3,3)\}$.

Function Comprehension A second way of defining a function is to express the "rule" that determines the elements it contains, without listing them explicitly. For example "the function that maps every number to itself" defines the identity function. Z provides two ways of doing this, one using standard set comprehension

notation, the other using "function construction" notation. See below.

Notation

The Z notation supports several ways of representing a function. A function extension expression may use ordered pair notation or maplet notation. The following example illustrates two ways to define the function that maps each integer between 1 and 3, inclusive, to its double.

$$f = \{(1,2),(2,4),(3,6)\} \tag{0.0.0.1}$$
$$= \{1 \mapsto 2, 2 \mapsto 4, 3 \mapsto 6\} \tag{0.0.0.2}$$

More generally:

$$\langle e_1, \ldots e_n \rangle = \{(1,e_1), \ldots (3,e_n)\} \tag{0.0.0.3}$$
$$= \{1 \mapsto e_1, \ldots 3 \mapsto e_n\} \tag{0.0.0.4}$$

Z also supports two ways to define a function intensionally – in terms of a property rather than an explicit list of elements.

Remark 1 *FIXME: set comprehension expression for functions:*

$$f = \{x, y \mid y = 2x \bullet (x,y)\} \tag{0.0.0.5}$$
$$= \{x, y \mid y = 2x \bullet x \mapsto y\} \tag{0.0.0.6}$$

Function construction notation...

$$f = \{\lambda \tag{0.0.0.7}$$

Sequences

Remark 2 *Lay down the basic concepts, terminology, and notation, for later use in discussing sampling, etc.*

Definition

Sequence Informally, a sequence is an ordered set. The Z specification says "A sequence is a particular form of function, where the domain elements are all the natural numbers from 1 to the length of the sequence."

Notation

The Z notation uses angle brackets to form a sequence expression:

$$\langle e_1, \ldots e_n \rangle = \{(1, e_1), \ldots (3, e_n)\} \tag{0.0.0.8}$$
$$= \{1 \mapsto e_1, \ldots 3 \mapsto e_n\} \tag{0.0.0.9}$$

Multisets

We follow Singh et al. [2007], with some modifications.

Definitions

Multiset

Element

Carrier The carrier of an mset is the set from which its elements is drawn.

Generator The generators of an mset are the elements of its carrier set.

Multiplcity The multiplicity of an element of an mset is the number of times it "occurs" ("appears", etc.).

Cardinality The cardinality (size) of an mset is the sum of the multiplicities of its elements. The cardinality of the carrier of an mset is the number of elements it contains.

Notation

The mset containing one a, two b, and three c

$$M = \{(a, 1), (b, 2), (c, 3)\} \tag{0.0.0.10}$$

can be written as follows:

Multiplicative notation The following are equivalent:

- $[[a, b, b, c, c, c]]$
- $[a, b, b, c, c, c]$
- $[a, b, c]_{1,2,3}$
- $[a^1, b^2, c^3]$
- $[a1, b2, c3]$

Note order is irrelevant.

Linear notation The following are equivalent:

- $[a] + 2[b] + 3[c]$
- $\{[a], 2[b], 3[c]\}$ Note that this style combines multiplicative notation for mset elements with standard set notation ($\{\ldots\}$) for the mset itself.

One advantage of the linear notation is that it allows us to have non-integral and non-positive multiplicities; for example, in $\{[a], -0.5[b], \pi[c]\}$ element a occurs once, b occurs -0.5 times, and c occurs π times.

In addition, linear notation allows a concise variant somewhat like a stem-and-leaf display:

$$\{[a], 2[b], 2[c], 3[d]\} = \{1[a], 2[b, c], 3[c]\}$$

Stem and Leaf

Multisets can be represented using stem-and-leaf tables:

Multisequences

A multiset is a set of ordered pairs, and is therefore unordered. Just as we can extend the concept of set to form a sequence, we can extend the notion of multiset to form a multisequence.

Multisequence A multisequence is a sequence of multiset elements.

We use the same angle bracket notation we use for set sequences:

$$\begin{aligned} \langle [a], 2[b], 3[c] \rangle &= \{(1, (a, 1)), (2, (b, 2)), (3, (3, c))\} && \text{(0.0.0.11)} \\ &= \{1 \mapsto (a, 1), 2 \mapsto (b, 2), 3 \mapsto (3, c)\} && \text{(0.0.0.12)} \\ &= \langle a, b, b, c, c, c \rangle && \text{(0.0.0.13)} \end{aligned}$$

Since multisets are unordered, we have

$$\{[a], 2[b], 3[c]\} = \{2[b], [a], 3[c]\} = \{3[c], 2[b], [a]\} = \ldots \quad \text{(0.0.0.14)}$$

Multisequences are ordered, so for example

$$\langle [a], 2[b], 3[c] \rangle \neq \langle 2[b], [a], 3[c] \rangle \quad \text{(0.0.0.15)}$$

since

$$\{(1, (a, 2), (2, (b, 2), (3, (c, 3)\} \neq \{(1, (b, 2)), (2, (a, 2)), (3, (c, 3))\} \quad \text{(0.0.0.16)}$$

alternatively

$$\{(1 \mapsto (a, 2), 2 \mapsto (b, 2), 3 \mapsto (c, 3)\} \neq \{(1 \mapsto (b, 2)), 2 \mapsto (a, 2)), 3 \mapsto (c, 3))\}$$
(0.0.0.17)

Also: $\langle a, b, b, c, c, c \rangle \neq [a, b, b, c, c, c]$

Series

Counting and Combining

4

What is combinatorics?

> Combinatorics can rightly be called the mathematics of counting. More specically, it is the mathematics of the enumeration, existence, construction, and optimization questions concerning nite sets. (Mazur, guided tour)

> The concept of configuration can be made mathematically precise by defining it as a mapping of a set of objects into a finite abstract set with a given structure; for example, a permutation of n objects is a bijection of the set of n objects into the ordered set 1, 2, ..., n. Nevertheless, one is only interested in mappings satisfying certain constraints.
>
> Just as arithmetic deals with integers (with the standard operations), algebra deals with operations in general, analysis deals with functions, geometry deals with rigid shapes, and topology deals with continuity, so does combinatorics deal with configurations. Combinatorics counts, enu- merates,* examines, and investigates the existence of configurations with certain specified properties.

[4] C Berge. *Principles of Combinatorics*. Academic Press, April 1971

Counting Principles

Remark 3 *Most of these are just restatements of basic arithmetic, expressed in terms of doing things. Why bother? I suspect that thinking in terms of sequences of actions makes it easier to do combinatorics. Also, these ideas only seem to be articulated as principles in elementary texts for e.g. high school algebra.*

Remark 4 *But isn't combinatorics just the science of counting?*

Principle 1 (Fundamental Principle of Counting)

Principle of addition : if there are a ways of doing one thing and b ways of doing another, and we cannot do both, then there are $a + b$ ways to choose one thing to do. This is just a restatement of a set-theoretic definition of addition in terms of union of sets.

In combinatoric texts something like this is more typical:

Addition Rule. If A and B are finite, disjoint sets, then $A \cup B$ is finite and $|\, A \cup B \,| = |\, A \,| + |\, B \,|$.

Principle of multiplication : if there are a ways of doing one thing and b ways of doing another, then there are $a \cdot b$ ways of doing both. This is a restatement of a set-theoretic definition of addition in terms of cartesian products.

Computability and Decideability

Remark 5 *Why important for intro to stats? Mainly for historical reasons; both effective procedure and axiom of choice emerged at roughly the same time. Also, an understanding of effective proc sharpens understanding of choice and randomness.*

Effective Procedure

Choice and Chance

The Axiom of Choice

The Axiom of Choice is of enormous important in mathematics generally. Statistics is no exception; the significance of the axiom will become especially apparent when we discuss the concept of random sample.

> The principle of set theory known as the Axiom of Choice has been hailed as probably the most interesting and, in spite of its late appearance, the most discussed axiom of mathematics, second only to Euclid's axiom of parallels which was introduced more than two thousand years ago (Fraenkel, Bar-Hillel & Levy 1973, ğII.4). The fulsomeness (*sic*) of this description might lead those unfamiliar with the axiom to expect it to be as startling as, say, the Principle of the Constancy of the Velocity of Light or the Heisenberg Uncertainty Principle. But in fact the Axiom of Choice as it is usually stated appears humdrum, even self-evident. For it amounts to nothing more than the claim that, given any collection of mutually disjoint nonempty sets, it is possible to assemble a new set a transversal or choice set containing exactly one element from each member of the given collection. Nevertheless, this seemingly innocuous principle has far-reaching mathematical consequences many indispensable, some startling and has come to figure prominently in discussions on the foundations of mathematics. It (or its equivalents) have been employed in countless mathematical papers, and a number of monographs have been exclusively devoted to it.[5]

[5] John L. Bell. The axiom of choice. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2013 edition, 2013. URL http://plato.stanford.edu/archives/win2013/entries/axiom-choice/

Often stated in terms of choice functions.
Variants:
AC1: Any collection of nonempty sets has a choice function."
AC2: Any indexed collection of sets has a choice function.
Or relations:

Axiom 1 (Axiom of Choice) *For every family $\mathcal{F}$ of nonempty disjoint sets there exists a selector, that is, a set S that intersects every $F \in \mathcal{F}$ in precisely one point.*[6]

[6] Krzysztof Ciesielski. *Set Theory for the Working Mathematician*. Number 39 in London Mathematical Society student texts. Cambridge University Press, Cambridge ; New York, 1997

Transversal: In a 1908 paper Zermelo introduced a modified form of AC. Let us call a transversal (or choice set) for a family of sets H

any subset T H for which each intersection T X for X H has exactly one element. As a very simple example, let H = 0, 1, 2, 3. Then H has the two transversals 0, 1, 2 and 0, 1, 3. A more substantial example is afforded by letting H be the collection of all lines in the Euclidean plane parallel to the x-axis. Then the set T of points on the y-axis is a transversal for H.

So we have choice functions and choice sets.

"Let us call Zermelo's 1908 formulation the combinatorial axiom of choice:

CAC: Any collection of mutually disjoint nonempty sets has a transversal." (bell)

The problem:

> It is to be noted that AC1 and CAC for finite collections of sets are both provable (by induction) in the usual set theories. But in the case of an infinite collection, even when each of its members is finite, the question of the existence of a choice function or a transversal is problematic[4]. For example, as already mentioned, it is easy to come up with a choice function for the collection of pairs of real numbers (simply choose the smaller element of each pair). But it is by no means obvious how to produce a choice function for the collection of pairs of arbitrary sets of real numbers.[7]

[7] John L. Bell. The axiom of choice. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2013 edition, 2013. URL http://plato.stanford.edu/archives/win2013/entries/axiom-choice/

Footnote:

> The difficulty here is amusingly illustrated by an anecdote due to Bertrand Russell. A millionaire possesses an infinite number of pairs of shoes, and an infinite number of pairs of socks. One day, in a fit of eccentricity, the millionaire summons his valet and asks him to select one shoe from each pair. When the valet, accustomed to receiving precise instructions, asks for details as to how to perform the selection, the millionaire suggests that the left shoe be chosen from each pair. Next day the millionaire proposes to the valet that he select one sock from each pair. When asked as to how this operation is to be carried out, the millionaire is at a loss for a reply, since, unlike shoes, there is no intrinsic way of distinguishing one sock of a pair from the other. In other words, the selection of the socks must be truly arbitrary.

The axiom of choice and probability (randomness) are different concepts. See http://math.stackexchange.com/questions/29381/picking-from-an-uncountable-set-axiom-of-choice

Chance and Randomness

See [8]

Indeterminacy, disorder, chaos, stochastic process, etc.

chance of a process, randomness of its product

Chance: physical; randomness: mathematical?

[8] Antony Eagle. Chance versus randomness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014. URL http://plato.stanford.edu/archives/spr2014/entries/chance-randomness/

"It is safest, therefore, to conclude that chance and randomness, while they overlap in many cases, are separate concepts."[9]

Process v. Product concepts

"Of course the terminology in common usage is somewhat slippery; it's not clear, for example, whether to count random sampling as a product notion, because of the connection with randomness, or as a process notion, because sampling is a process."

> The upshot of this discussion is that chance is a process notion, rather than being entirely determined by features of the outcome to which the surface grammar of chance ascriptions assigns the chance. For if there can be a single-case chance of ¡ for a coin to land heads on a toss even if there is only one actual toss, and it lands tails, then surely the chance cannot be fixed by properties of the outcome lands heads, as that outcome does not exist.[2] The chance must rather grounded in features of the process that can produce the outcome: the coin-tossing trial, including the mass distribution of the coin and the details of how it is tossed, in this case, plus the background conditions and laws that govern the trial. Whether or not an event happens by chance is a feature of the process that produced it, not the event itself.[10]

> a process conception of randomness makes nonsense of some obvious uses of random to characterise an entire collection of outcomes of a given repeated process. This is the sense in which a random sample is random: it is an unbiased representation of the population from which it is drawnand that is a property of the entire sample, not each individual member. While many random samples will be drawn using a random process, they need not be....To be sure that our sample is random, we may wish to use random numbers to decide whether to include a given individual in the sample; to that end, large tables of random digits have been produced, displaying no order or pattern (RAND Corporation 1955). This other conception of randomness, as attaching primarily to collections of outcomes, has been termed product randomness.(eagle)

> If the actual process that generate the sequences are perfectly deterministic, it may be that a typical product of that process is not random. But we are rather concerned to characterise which of all the possible sequences produced by any process whatsoever are random, and it seems clear that most of the ways an infinite sequence might be produced, and hence most of the sequences so produced, will be random.(eagle)

...concentrate on the sequence of outcomes as independently given mathematical entities, rather than as the products of a large number of independent Bernoulli trials...

Goal: devise mathematics to "capture the intuitive notion of randomness."

Compare logicians' attempts to capture the intuitive notion of logical consequence.

[9] Antony Eagle. Chance versus randomness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014. URL http://plato.stanford.edu/archives/spr2014/entries/chance-randomness/

[10] Antony Eagle. Chance versus randomness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014. URL http://plato.stanford.edu/archives/spr2014/entries/chance-randomness/

Howson and Urbach (1993: 324) that it seems highly doubtful that there is anything like a unique notion of randomness there to be explicated.

Intuitions about randomness:

- a property of a sequence
- indeterminism
- epistemic randomness

See also https://www.cs.auckland.ac.nz/ chaitin/sciamer.html

Infinity and the Limit Concept

Measure and Integration

Remark 6 *Why this? Mainly just as a notational convenience. Even if we target an audience with minimal math, it is useful to have* $\int$ *(or* $\overset{+}{\lambda}$*) to indicate the area under a curve. And that is critical to the concept of probability of a continuous random variable, which is one of the main concepts we want to get across.*

Furthermore the basic concepts are not that difficult: differentiation as the limit of a ratio of differences, integration is the limit of a sum of products. The details of how one might actually do this may be daunting for many readers, but the basic concepts are quite simple and intuitive. Especially with lambda notation.

Remark 7 *What about differentiation?*

Thinking Exponentially (and Logarithmically)

Remark 8 *Task 1: convince reader that exp and log, power and root, are special. Task 2: Provide intuitive, clear, simple explanation. Task three: show relation to stats thinking.*

Learning to think in terms of exponents and logarithms is one of the keys to learning to think statistically. Fortunately it's not too difficult, and it happens to be fascinating.

Getting from here to there. We normally think linearly. The number line is treated as additive; to get from a to b on the number line, you add the (possibly negative) difference between the two: $a + (a - b) = b$.

Exponentiation offers a different way of thinking about the relation between two numbers. Instead of viewing their difference as spatial distance $(a - b)$, we think of it in terms of difference in growth. So the difference between, say, 3 and 9 is not $6(=\mid 3 - 9 \mid)$, but $2(= log_3\ 9)$. Here we take 2 as the number of (natural) growth cycles it takes for 3 to turn into 9: $3^2 = 9$.

Why *growth*? Because exponentiation is self-contained, so to speak. Getting from 3 to 9 linearly involves adding something external (6) to 3. But 3^2 does not involve any external "factor" in this way (other than the multiplication operation); the difference between 3 and 9 is construed in terms of 3 alone; "raising 3 to a power" is construed as an operation involving 3 alone. The task is to find how many times this something must be done for 3 to "turn into" (rather than "arrive at") 9.

Traditional terminology (now archaic) captured this; as late as the 19th century, a common term for exponentiation was "involution", meaning something like "turning in to itself".

Note that conventional terminology, being derived from Greek geometry (3^2 means "three *squared*"), is misleading. Exponentiation has nothing essential to do with geometry. (And note that the Greeks did not have a genuine concept of exponentiation; they never came up with the kind of algebraic thinking involved in finding powers and roots.) The Arabic-speaking mathematician who is credited with

inventing algebra in something like the form we know it today (Al-Khawarizmi) did not use the term "squared"; his (Arabic) word for what we call a squared quantity was *mâl*, meaning "cattle, stock, wealth". Furthermore, the Arabic term for multiplication was also conceptually distinct from the notion of repeated "folding" ("multiply" comes from the latin *multiplicare*, combining *multi* "many times" and *plicare*, from *plex* "fold"). The Arabic term was *darb*, "strike", and to form a *mâl*, one "strikes" a number *in itself*. The origins of this usage are not known, but note that the same term is used in minting ("strike a coin") and husbandry ("strike" was used to refer to copulation of e.g. livestock). So historically, exponentiation was did not emerge from either arithmetic nor geometry; rather, it emerged as a distinctive concept.

But exponentiation is also conceptually distinct even (or especially) in modern mathematics. A satisfactory definition of exponentiation requires some fairly advanced calculus; the simple concept of something multiplied by itself is not sufficient.

It even shows up in very practical matters. Calculation of compound interest turns out to be intimately related to exponentiation, for example.

Remark 9 *Story of e as discovery of a mathematically satisfying account of exponentiation. Exponentiation defined in terms of logarithms, rather than the other way around.*

The critical point of all this is that we should think of exponentiation as a special kind of operation, rather than as something derived from arithmetic (multiplication). Fortunately this is not particularly difficult, but it does require a basic change in perspective.

The payoff will come when we consider probability distributions, in which exponentiation plays a critical role.

Remark 10 *Also logarithmic scales, logit, etc. Lots of places where logs and exponents are critical.*

Remark 11 *Exponentiation as involution. Inflection points. Symmetry. Constructing the numbers from exponentiation* ($\lim_{x\to 0} e^x = 0$).

Remark 12 *Powers and roots v. exponentiation. Difference between* $f(x) = x^a$ *(powers) and* $f(x) = a^x$ *(exponentiation).*

Sequences: any term of the following sequences can be used to form a function, e.g. $f(x) = x^2, g(x) = 2^x$.

$x^0, x^1, x^2, .., x^e, \ldots, x^n$ Power (geometric) sequence (0.0.0.18)

$0^x, 1^x, 2^x, .., e^x, \ldots, n^x$ Exponential sequence (0.0.0.19)

Powers and Roots

Remark 13 *Powers and roots as "phases" ("poles", polarity?) of exponentiation, with* 1 *as the "inflection point". But we need a different term, "inflection point" should be reserved for changes in the derivative of a curve. We want something that indicates a phase shift, where powers switch to roots and vice-versa. "Phase" by analogy to phases of matter (solid, liquid, gas), not of cycles (waves). I.e. qualitative change, not cyclic orientation. Critical point? (Freezing point, melting point, etc.)*

This can best be explained by example.

Let's review some basic rules:

$$a^b = a \cdot a \cdots a \quad \text{(b times)} \tag{0.0.0.20}$$

$$a^{-b} = \frac{1}{b} \tag{0.0.0.21}$$

$$a^{1/b} = \sqrt[b]{a} \tag{0.0.0.22}$$

$$a^{b/c} = \sqrt[c]{a^b} \tag{0.0.0.23}$$

First powers. Conventionally, the expression a^b tells us to multiply a by itself b times (that is, $b - 1$ multipication operations involving b terms 1). This makes perfect sense, if b is a whole number and $b > 1$. But what about, say, $a^{1.5}$? Or even worse, a^π? As it happens, it is fairly easy (but perhaps not trivial) to define a^b for all rational b (like 1.5), but defining it for irrational b (like π) is another matter. Explaining how this is done is beyond the scope of this paper, so for our purposes let's just assume that a^b is defined for all real b.

So $a^{1.5}$ and a^π are defined, but what does it mean to multiple a by itself 1.5 or π times? Not very intuitive. One strike against the conventional explanation of exponentiation in terms of multiplication.

For now, though, let's stick with the multiplication idea. The point here is that we can treat a^b as a problem to be solved, or a task to be accomplished: find c such that $a^b = c$. Let's look at the structure of the task. We are given two operands, a and b, and an operation

(exponentiation); the task is to find out where these lead us. In other words, we do not know where we are going to end up, but we know how to get there: by raising a to the bth power. In brief, our task is to reach the destination, given the means of getting there.

Now consider roots. The expression $\sqrt[b]{a}$, unlike a^b, tells us where we are to end up, but does not tell us how to get there. "Find the bth power of a means "perform the exponent operation b times on a; it tells us what to do; but "find the square root of a" does not mean "perform the square root operation on a"; it tells us what the destination is, but not what we need to do to get there. In other words, $\sqrt[b]{a}$ denotes a value related to a and b, not an operation that uses a and b; in contrast, a^b, although it indirectly denotes the destination value, directly denotes the operation to use to get there.

Remark 14 *Improve this. Both can be viewed as denoting either a value, a process, a device, or all of the above. Why not think of $\sqrt[b]{a}$ as denoting an operation? Mainly because there is no such operation, or at least it isn't normally understood in that way. Clearly in contrast to a^b.*

Remark 15 *Correction: a^b only seems to tell us what to do, because it does so in the case of integer exponents. But in the case of real exponents, it precisely does not tell us what to do. But this is evidence of the specialness of exponentiation. The arithmetic operators do tell us what to do; that is how they are defined. But we cannot define exp and root in this way. In fact the best we can do is find a procedure for approximating the value.*

Alternatively: arithmetic ops defined via effective procedures. Exp ops defined in terms of meaning, not procedure. There may be many distinct procedures that can be used to solve them.

This is obvious in the case of square root. Easy to define, but I would guess even relatively well-educated people would not know how to go about computing a square root by hand. There are dozens of methods for computing square roots, all of which (?) only approximate the answer.

Ditto for the trig funcs. Easy to provide a geometric definition, hard to say how to compute. Well, semi-hard, not as hard as exp and log.

Remark 16 *Put this in a graph: Here's another symmetry: given $b > 1$, as b increases, a^b decreases if $0 <\mid a \mid< 1$, but a^b increases if $\mid a \mid> 1$; if $a = 0$, then $a^b = 0$ for all b, and if $a = 1$, then $a^b = 1$ for all b. So here $a = 1$ is a critical point where a^b changes direction, so to speak, and $a = 0$ is a critical point - call it a "constant point" where a^b is constant.*

Exponentiation and Co-exponentiation

Random Variables

Random selection devices

Functions

First- and Second-order Functions

Remark 17 *Family of functions/curves: function and meta-function. $y = e^x$ is a function; $y = e^{\alpha\ x}$ is a meta-function. Better: second-order function.*

Use lambda abstraction to demonstrate the difference: partial application of a second-order function yields a first-order function.

Terminology: apply higher-order functions to parameters, first-order functions to arguments.

Equations, Graphs, and Curves

Remark 18 *Ideally, one would be able to instantly visualize the curve upon seeing the equation. But there are many many functions for which this is not so easy. But with a little practice it becomes relatively easy to know the basic shape of a curve from a glance at its equation. So one purpose here is training in the art of seeing the curve in the equation. Example: once you understand the relation between functions of e and their curves, such as the curves of e^x and e^{-x}, then it becomes relatively easy to see what shape the curve of the Gaussian PDF ought to have.*

Functions of the form e^n are particularly common, where n itself can be any sort of expression, e.g. $e^{-(x/\ \lambda)^k}$ (Weibull pdf).

Mastering the shape of an equation really means mastering the shapes of a family of equations.

We can show quite clearly what shape a family of functions has by using animation to show what happens as the parameters vary. This will both expose the general shape of the (meta) function, and the role of the parameters.

Remark 19 *A second critical point is that the parameters of a meta-function*

Characteristics of curves

Kurtosis, skew, scedasticity - fancy Greek terms for sharpness, skew, and scatter. But useful for classifying curves by shape.

Thin/fat tails. http://en.wikipedia.org/wiki/Fat-tailed_distribution. E.g. Cauchy distrib, stable distribs (except the normal)

Part II

Philosophy

Remark 20 *Relevance of philosophy to stats? Bridge between mathematics and world.*

Radical empricism: data first, then theory.
Rationalism: theory, then data.

Causality

Remark 21 *Factor analysis - factor = causal factor*

Part III

Science

Science originally "natural philosophy".

Remark 22 *Science and scientism, pseudo-science, cargo-cult science.*

Remark 23 *Scientific method - ties philosophy to experience.*

Remark 24 *On the relationship between mathematics and the world. Involving fundamental philosophical ideas, also pragmatics of doing "science".*

The fundamental issue is what sorts of claims can statistics make about the world, how should we take them, etc. The goal here is of course clarity.

This is the natural place for a historical perspective, as well. How did statistical practice (and theory) evolve?

Remark 25 *Stress: empiricism as a philosophy did not pan out - we are not mere observers of given data.*

Paradigmatic cases? Astronomical observations; invention/discovery of quantifiable temperature, psychometrics, etc.

Measurement and Error

Remark 26 *The previous section discuss the mathematical concept of measure. This section discusses theories of empirical measurement.*

Part IV

Probability

Probability

Measure Theory

Probability Measures

Joint Probabilities

Conditional Probabilities

Likelihood

"In statistics, a likelihood function (often simply the likelihood) is a function of the parameters of a statistical model. The likelihood of a set of parameter values, , given outcomes x, is equal to the probability of those observed outcomes given those parameter values, that is $\mathcal{L}(\theta \mid x) = P(x \mid \theta)$ (http://en.wikipedia.org/wiki/Likelihood_function)

Random Variables

Discrete Rvs

Continuous Rvs

Algebra of Rvs

Moments

"In probability theory and statistics, a central moment is a moment of a probability distribution of a random variable about the random variable's mean; that is, it is the expected value of a specified integer power of the deviation of the random variable from the mean. The various moments form one set of values by which the properties of a probability distribution can be usefully characterised. Central moments are used in preference to ordinary moments, computed in terms of deviations from the mean instead of from the zero, because the higher-order central moments relate only to the spread and shape of the distribution, rather than also to its location."http://en.wikipedia.org/wiki/Moment_about_the_mean

Remark 27 *NB analogy between moment of a distrib (a function) and derivative of a function. Nth moment, Nth derivative, etc.*

Moment: expected value of nth involution of deviation (so we could generalize, so $y = log_{\bar{X}-\mu} m$ *for any m, yielding m as the* y^{th} *moment.*

Compare finding nth derivative, or, given a function, finding the nth anti-derivative (integral).

So why moments? I suspect they're like derivatives: just as derivatives tell us something about the original function, moments tell us something about the original distribution (i.e. the random var, which is a function). Or so I would expect, based on nothing more than a principle of symmetry.

Expected Value

Variance

Skew

Probability Distributions

http://en.wikipedia.org/wiki/Probability_distribution

Remark 28 *Stress: connection between concepts of random var and prob. distrib*

A probability distribution is just a function, or rather a family of functions expressed as parameterized equations.

Remark 29 *Family of functions is the critical idea. Goal is to pick the best family, then the best function from the family.*

Ways of specifying probability distributions:

Probability Mass Function

Probability Density Function

Cumulative Distribution Function

Characteristic Function

Moment-generating Function

"In probability theory and statistics, the moment-generating function of a random variable is an alternative specification of its probability distribution. Thus, it provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions. There are particularly simple results for the moment-generating functions of distributions defined by the weighted sums of random variables. Note, however, that not all random variables have moment-generating functions."http://en.wikipedia.org/wiki/Moment-generating_function

Hazard Function

I.e. failure rate.

"Failure rate is the frequency with which an engineered system or component fails, expressed, for example, in failures per hour. It is often denoted by the Greek letter λ (lambda) and is important in reliability engineering." http://en.wikipedia.org/wiki/Hazard_function#hazard_function

"Calculating the failure rate for ever smaller intervals of time, results in the hazard function (also called hazard rate), $h(t)$. This becomes the instantaneous failure rate as t tends to zero...A continuous failure rate depends on the existence of a failure distribution, $F(t)$, which is a cumulative distribution function that describes the probability of failure (at least) up to and including time t..." http://en.wikipedia.org/wiki/Hazard_function#hazard_function

See Bathtub curve

See Weibull distrib

Well-known Distributions

Discrete

Degenerate Distrib

"The degenerate distribution at xo, where X is certain to take the value xo. This does not look random, but it satisfies the definition of random variable. This is useful because it puts deterministic variables and random variables in the same formalism." http://en.wikipedia.org/wiki/Degenerate_distribution

Continuous

Gauss

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{0.0.0.24}$$

But:

$$e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{e^{\frac{(x-\mu)^2}{2\sigma^2}}} = \frac{1}{\sqrt{e^{\frac{(x-\mu)^2}{\sigma^2}}}} = \frac{1}{\sqrt{e^{\frac{x-\mu}{\sigma}}}} \tag{0.0.0.25}$$

Now set $\mu = 0, \sigma = 1$. Then

$$\frac{1}{\sigma\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}} = 0.398942 \tag{0.0.0.26}$$

and

$$e^{-\frac{(x-\mu)^2}{2\sigma^2}} = e^{-\frac{x^2}{2}} = \frac{1}{e^{\frac{x^2}{2}}} \tag{0.0.0.27}$$

Then if x=o, we have

$$e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{e^{\frac{0}{2}}} = \frac{1}{e^0} = \frac{1}{1} = 1 \tag{0.0.0.28}$$

so when $\mu = 0, \sigma = 1, x = 0$,

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} = 0.398942 \tag{0.0.0.29}$$

As x grows in either direction from zero, $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ ...

The turning point is set by σ^2, since that determines when $\frac{(x-\mu)^2}{2\sigma^2}$ pivots about 1. As x gets larger beyond that point, $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ gets smaller.

Main point: $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ never exceeds 1, since the exponent is never less than zero, and $e^0 = 1$. Then for all $x > 0$, e^x gets bigger, so $\frac{1}{e^x}$ will always be < 1. So the product of the two factors will always be less than 0.398942... The bell shape comes from the varying rate of change of $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, which reflects the growth curve of $\frac{1}{e^x}$, which declines relatively rapidly near zero, but flattens out as x increases.

The key is in the derivatives:

$$\overline{\lambda}x.e^x = e^x \tag{0.0.0.30}$$

$$\overline{\lambda}x.e^{-x} = -e^{-x} \tag{0.0.0.31}$$

$$\overline{\lambda}x.x^{\sigma x} = \sigma x^{\sigma x} \tag{0.0.0.32}$$

$$\overline{\lambda}x.x^{-\sigma x} = -\sigma x^{-\sigma x} \tag{0.0.0.33}$$

So as σ varies, so does the flattness of the curve.

Remark 30 *Inflection points: in addition to the max and min, we have the points at which the second derivative switches from pos. to neg. In other words, the point at which the normal curve starts to flatten. This happens when $x = \mu \pm \sigma$. This makes intuitive sense since it is the point at which $\frac{(x-\mu)^2}{2\sigma^2}$ is 1 (assuming $\mu = 0, \sigma = 1$), which makes it the place where the exponent switches from power to root.*

Remark 31 *Plot e^{nx} as n varies. As n decreases, the graph becomes flatter; as it increases, it begins to approximate a sharp L shape. This is the effect of varying σ; as it decreases, the exponent increases. Better: plot $e^{\frac{x}{n}}$.*

All of which goes to show, that the normal curve is a kind of exponential curve, which makes it a kind of growth curve. It describes the way error grows as more observations are made.

Sampling

Remark 32 *Distinguish between mathematical and empirical usage. Here we talk of carrier sets instead of populations, elements instead of sample units, values instead of measurements, etc.*

If we think of sampling clearly, in terms of sets, sequences, etc. then the inferential stuff becomes intuitively clear.

Samples

A sample is a sequence of elements drawn from the carrier set (population).

Sampling Continuous Random Vars

Discrete v. continuous sampling? A continuous sample would have to select ranges rather than elements. But often we want discrete sampling of a continuous var; e.g. distance from bullseye. This would give us a multiset, but what would the elements and multiplicities be? E.g. 3[.25] would mean three events of striking within .25 of the center.

Sampling Sequences

Remark 33 *Or: sample sequences. Sequence stuff from I, from the perspective of sampling.*

The Sampling Pool

Or sample pool

The sampling pool is the union of the samples in the sampling sequence.

Discrete

Remark 34 *What we're going to want is counts of number of possible samples, number of unique samples (combinations), distribution of samples when the sampling sequence is long, distribution of elements in the sampling pool. Etc.*

Multivariate stuff

Part V

Statistics

Remark 35 *What distinguishes stats from probability? Probability seems to have everything stats has: expected value (for mean), etc. Even statistical inference is based entirely on a result from probability, the Central Limit Theorem. So why not call it all probability?*

Statistics measures something else - location (central tendency), spread (deviations), etc.

Descriptive Statistics

Location

Spread

Curve fitting

aka Hypothesis Testing

Association

Inferential Statistics

From Sample to Population

From Correlation and Causality

From Manifest to Occult

IOW, from observable to latent variables.

EDA

Exploratory Data Analysis

Tukey: http://pi.lib.uchicago.edu/1001/cat/bib/151262

Hartwig: http://pi.lib.uchicago.edu/1001/cat/bib/9149154

Part VI

Causality

Factor Analysis

Part VII

Appendices

Appendices

Bibliography

Bibliography

ISO/IEC 13568:2002. *Information technology – Z formal specification notation – Syntax, type system and semantics*. ISO, Geneva, 2002. URL http://www.iso.org/iso/catalogue_detail.htm?csnumber=21573.

John L. Bell. The axiom of choice. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2013 edition, 2013. URL http://plato.stanford.edu/archives/win2013/entries/axiom-choice/.

C Berge. *Principles of Combinatorics*. Academic Press, April 1971.

Krzysztof Ciesielski. *Set Theory for the Working Mathematician*. Number 39 in London Mathematical Society student texts. Cambridge University Press, Cambridge ; New York, 1997.

Antony Eagle. Chance versus randomness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014. URL http://plato.stanford.edu/archives/spr2014/entries/chance-randomness/.

D. Singh, A. M. Ibrahim, T. Yohanna, and J. N. Singh. An overview of the applications of multisets. *Novi Sad J. Math.*, 37(2):73–92, 2007.

Stephen M Stigler. *The History of Statistics: the Measurement of Uncertainty Before 1900*. Belknap Press of Harvard University Press, Cambridge, Mass., 1986.