G. A. REYNOLDS

THINK STATS

A CONCEPTUAL INTRODUCTION TO STATISTICS

FOR THE SKEPTICAL, THE PESSIMISTIC, AND THE MILDLY DISTURBED

Remark 1 Two tasks: conceptual and computational. Understand stats first, then learn how to *do* stats using software (statistics machines).

Contents

Introduction

What is statistics[1]?

> 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.
>
> **?**

> 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.
>
> **?**

[1] test note

Part I

Mathematics & Logic

Sets ‘n Stuff

We use the Z Specification notation [2].

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]

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

[2]

[3]

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.

$$\begin{aligned} f &= \{(1,2),(2,4),(3,6)\} && (0.0.0.1)\\ &= \{1 \mapsto 2, 2 \mapsto 4, 3 \mapsto 6\} && (0.0.0.2) \end{aligned}$$

More generally:

$$\begin{aligned} \langle e_1, \ldots e_n \rangle &= \{(1,e_1), \ldots (3,e_n)\} && (0.0.0.3)\\ &= \{1 \mapsto e_1, \ldots 3 \mapsto e_n\} && (0.0.0.4) \end{aligned}$$

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

Remark 2 FIXME: set comprehension expression for functions:

$$\begin{aligned} f &= \{x, y \mid y = 2x \bullet (x,y)\} && (0.0.0.5)\\ &= \{x, y \mid y = 2x \bullet x \mapsto y\} && (0.0.0.6) \end{aligned}$$

Function construction notation...

$$f = \{\lambda \qquad (0.0.0.7)$$

Sequences

Remark 3 Lay down the basic concepts, terminology, and notation, for later use in discussing sampling, etc.

cf. Series

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:

$$\begin{aligned}\langle e_1, \ldots e_n\rangle &= \{(1, e_1), \ldots (3, e_n)\} && (0.0.0.8)\\ &= \{1 \mapsto e_1, \ldots 3 \mapsto e_n\} && (0.0.0.9)\end{aligned}$$

Multisets

We follow **?**, 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

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

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:

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

Since multisets are unordered, we have

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

Multisequences are ordered, so for example

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

since

$$\{(1, (a, 2), (2, (b, 2), (3, (c, 3)\} \neq \{(1, (b, 2)), (2, (a, 2)), (3, (c, 3))\}$$
(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: $\langle a,b,b,c,c,c\rangle \neq [a,b,b,c,c,c]$

Counting and Combining

4 4

What is combinatorics?

Combinatorics can rightly be called the mathematics of counting. More specifically, it is the mathematics of the enumeration, existence, construction, and optimization questions concerning finite 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.

Counting Principles

Remark 4 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 5 But isn't combinatorics just the science of counting?

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 $\mid A \cup B \mid = \mid A \mid + \mid B \mid$.

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.

Progressions and Means

Arithmetic

Definition 1 (Arithmetic Progression)

Definition 2 (Arithmetic Mean)

$$\frac{x_1 + \cdots x_n}{n} = \frac{\sum_1^n x_n}{n} = \frac{\sum x}{n}$$

Geometric

Definition 3 (Geometric Progression)

Definition 4 (Geometric Mean)

$$\sqrt[n]{x_1 \times \cdots \times x_n} = \sqrt[n]{\Pi_1^n x} = \sqrt[n]{\Pi x}$$

Harmonic

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]

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]

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

[5]

[6]

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]

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?

[7]

[8]

"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]

[10]

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

The Function Ladder: Differentiation 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 Furthermore there is an aesthetic component: symmetry. Differentiation and integration as relations among functions; comparable to the relation between a function and its inverse.

Pedagology usually presents differentiation in terms of physical motion, instantaneous rate of change, etc. Integration is presented as a matter of finding area under a curve. But neither of these really capture what's at stake; this is obvious if you ask what the area under a curve has to do with probability. Physical metaphors are not necessary to acquire a solid intuitions of differentiation and integration. A (perhaps) better approach is to focus on the symmetrical relations involved. Differentiation and integration as higher-order functions that map other functions to their "natural" co-functions or counterparts. So we start by just asking, e.g. what is the relation between x^2 and $2x$, or x^3 and $3x^2$. We can easily see a pattern involving coefficients and exponents; the derivatives and integrals account for these.

Avoid "rate of change" talk at all costs. Mathematics is static; there is no change. So if $2x$ is the derivative of x^2 we need a way to express the intuition involved without appealing to a concept of change or motion. Just: ratio, a ratio function of fixed form but different value at each x. Not rate of change, but ratio of "sizes". We can change our focus of attention to different values of x and y, but the functions themselves do not do this; they are fixed, unmoving, rigid. So its like looking at a building and comparing its height to its width.

But let's face it: there's no way to eliminate intuitions of rate-of-change. Going from ratio of magnitudes to rate of change is natural.

[An alternative perspective: (organic) growth. A function grows; its derivative says how fast it grows, and its integral says how much "mass" it accumulates.]

[Another alternative: behavior. Functions "behave" differently at different places. etc.]

If $f(x)$ tells us where we are - how far along toward the end - then $f'(x)$ tells us how fast we're moving. Reverse perspective, and if $f(x)$ tells us where we are, then $\int f(x)$ tells us how much ground we've covered from the beginning. The former is about rate, the latter about accumulation.

The beauty of it is the symmetry. If we have f and g, where f is the derivative of g, then g is the integral of f. By looking at f we can see how fast g is moving; by looking at g, we can see how much ground f has covered. The location of f describes the rate of g; the location of g describes the accumulation of f.

If we think of f and g as two distinct moving bodies, then this fundamental relation between differentiation of the one and integration of the other - call it the conversion relation, by analogy with inverse relation - links their rates. They move at different rates and thus cover different distances over the same "time", but those rates and distances are related. Position, rate, and cumulative distance are the three things related by this conversion relation. Corresponding to three kinds of question: where is it? how fast is it moving? and how far has it traveled? The function itself answers the first question, its derivative answers the second, and its integral answers the third.

Note that there are two symmetries here. One a kind of inverse relation between two functions, derivative and anti-derivative, and the other a kind of transitive relation between a function, its derivative and its anti-derivative.

This exposes a fundamental difference between the notion of inverse function and the more special kind of inverse relation involved in differentiation and integration. The plain old inverse relation only involves two functions; there is no transitive relation to some third function. In other words, the inverse relationship is strictly symmetric but not transitive, and the "conversion" relation is both. You can go up and down the ladder. But the ladder is well-founded; you can go up (integrate) indefinitely, but you always hit bottom going down. More concretely: rates of rates of rates ... eventually end up at zero: no change at the nth degree rate. But accumulation of accumulation of accumulation ... just keeps going. (Since each "rung" in the upward ladder covers some ground.)

Example: there are infinitely many “rungs” in the integration ladder of $f(x) = x$, but going the other way always eventually bottoms out in zero.

Bottom line: fundamental theorem of calculus expresses a transitive relation.

But this means that each (well-behaved) function is actually, and essentially, part of a chain of functions. The relation between a function, its derivative, and its integral is intrinsic. So we cannot treat such functions as isolated individuals; rather they are akin to points on a continuum.

Compare this to the exponential ladder. Taking the nth derivative (integral) akin to taking the nth power (root). Compare moments about the mean.

Recursion?

Fun, but how relevant to stats? Maybe it will give us intuitions about distribs, etc. Derivatives tell us something about their original functions, etc. Moments about the mean tell us something about the original population. etc. Mathematical techniques of using derived things to reveal information about basic things.

Measure

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. $f(x) = x^2, g(x) = 2^x$.

$$0^x, 1^x, 2^x, .., e^x, \ldots, n^x \quad \text{Power sequence?} x^0, x^1, x^2, .., x^e, \ldots, x^n \quad \text{Exponential (geometric) sequence} \qquad (0.0.0.18)$$

$$(0.0.0.19)$$

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:

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

$$a^{-b} = \frac{1}{b} \qquad (0.0.0.21)$$

$$a^{1/b} = \sqrt[b]{a} \qquad (0.0.0.22)$$

$$a^{b/c} = \sqrt[c]{a^b} \qquad (0.0.0.23)$$

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 III

Philosophy

Remark 21 Relevance of philosophy to stats? Bridge between mathematics and world.

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

Causality

Remark 22 Factor analysis - factor = causal factor

Part IV

Science

Science originally “natural philosophy”.

Remark 23 Science and scientism, pseudo-science, cargo-cult science.

Remark 24 Scientific method - ties philosophy to experience.

Remark 25 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 26 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 27 The previous section discuss the mathematical concept of measure. This section discusses theories of empirical measure*ment*.

Part VI

Statistics

Remark 36 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

Mean

Ed. note 0.0.1 *key terms: uniform distrib; equity; equitable distrib; allocation; parameterization of r.v.; nth degreee equation/function/mean/deviation; moments as n^{th} mean/deviation.*

Interpretation of the Arithmetic Mean

Ed. note 0.0.2 $\frac{\sum x}{n-1}$ *is not the arithmetic mean; but it is a* kind *of mean. We could call it "mean of randomization", "free mean", or the like.*

We can treat the arithmetic mean as a description of a single particular distribution (set). It can be interpreted in multiple ways, one per polynomial degree. At degree zero, it is the value of the constant function; at degree 1, the midpoint of both the domain and range of a linear function; etc.

By contrast, we think of the "free mean" not as a determinate description, but as a kind of generalized representation of a family *of samples (sets), all of which are "equivalent up to" cardinality and and sum. "Equivalent up to cardinality and sum" is a fancy way of saying that each sample in the family has the same number of elements, which sum to the same total. It follows that each sample has the same mean. We could say "equivalent up to mean", except same mean does not imply same cardinality.*

In other words, given a specific sample set, there is an infinite number of different sample sets that "look the same" as the original set in *that they have the same cardinality and the same sum (and thus the same mean). The "free mean" describes that set; we can use it, so to speak, to generate specific samples, by random selection parameterized by the free mean.*

Another constraint: not only do we want our $n-1$ choices to be random; we also want them to have the "same amount" of randomness. In other words, we want our randomization devices to be identical (cf. i.i.d.). This maximizes total randomness. Note that we could "bias" the randomness by giving each device a different range.

Note that there is a basic tradeoff here. In order to maximize total randomness across $n-1$ devices, we are forced to restrict the range allocated to the devices. If we were to allocate more than the free mean amount to the devices, then the possibility arises that the total could be "used up" before we get through all $n-1$ devices. In that case we would not have $n-1$ free (random) choices. So we compromise, by arranging things such that even if every choice selects the maximum possible value, there will always be "enough" choice left for the remaining devices, up to $n-1$. In this extreme scenario, the first $n-1$ choices exhaust the entire total available, leaving zero for the final (n^{th}) element. But that's ok, because that element is deterministic in every case; it is never freely chosen.

So our compromise means that no value of our sample will exceed $\frac{\sum x}{n-1}$. This effectively means we limit the possible variance in our sample in order to guarantee maximal total randomness.

Note that the family (=set) of samples determined by the free mean is not exhaustive; there is yet another infinity of equivalent samples that are not "explainable" by the free mean. For example, consider the sample whose first element consumes the entire total, so that all remaining points get zero. This counts as equivalent up to cardinality and sum, but it has only one degree of freedom. Once the first element exhausts the available quantity, the remaining points are completely determined: they get zero, and no random choice is involved.

This suggests we should add something like "degree of randomness" to our notion of "equivalent up to". Samples determined by the free mean are equivalent to the original sample up to cardinality, sum, and (degree of) randomness. But this doesn't quite work, since randomness is not a property of specific samples, but of the selection process. And besides, it's not quite true that they are equivalent in this manner; all n points of the original were randomly selected, but only $n-1$ points in the equivalent sample were.

So perhaps a better terminology is: the "free mean" determines an infinite set of samples that are equivalent to the original "up to *cardinality and sum", and maximizes total randomness.*

Critical point: the notion that the free mean determines samples is based on interpreting it as a parameter setting the range of random variables. So the analogy is to a machine consisting of an array or chain of randomization devices whose output range is determined by a parameter setting, and the free mean serves as that parameter. "Running" the machine yields a maximally random sample equivalent up to the cardinality and sum to the original sample.

Ed. note 0.0.3 *Begin with distinction between determinate and indeterminate values (function as deterministic "process" v. random process). In our original set of values, each value construed as randomly generated. That's the critical difference between the statistical interpretation of arithmetic mean and its ordinary mathematical definition.*

Ed. note 0.0.4 *Something about concept (and history) of "normal distribution of error" and how it relates to concept of r.v. Isn't it the motivation for the various interpretations of mean and variation described below? We want to know which probability distrib provides the best "model" (= "explanation"?) of the data. NB: prob. dist. is a function, serves as deterministic mathematical model of randomness?*

Ed. note 0.0.5 *We start by "selecting" n random numbers; once they are selected, we find their sum* τ*. The goal is to "mimic" the original selection - make the same number of selections that sum to the same total. This would give us two distinct distributions with the same mean—two sets that are equivalent "up to the mean". Ideally every number in the second set would be selected at random, just as was the case in the first set. But there is no way to guarantee that n numbers selected at random will sum to a particular number. So we know from the beginning that at least one of our selections must not be random. Specifically, the last selection is determined by the sum of the preceding selections. This follows from the equation* $x_1 + \cdots + x_n = \tau \Rightarrow x_n = \tau - x_1 + \cdots x_{n-1}$*. In other words,* x_n *represents the quantity remaining to be allocated* *after the first $n-1$ numbers have been selected randomly, and it is completed determined: it is just the difference between τ and the sum $x_1 + \cdots x_{n-1}$.*

And so forth. The key point is that the goal of "mimicking" the original distribution [sidenote"Mimickery" may not be the right metaphor. What we're trying to do is find members of the equivalence class determined by the mean μ; that is, distributions whose mean is just μ.] involves two sub-goals that are in tension: maximizing randomness, and summing to the predetermined total. The task is to find the optimal combination of random and determinate values. That solution is in terms of "degrees of freedom", which is the designated terminology for "number of random choices involved". For a set of n values summing to τ, only $n-1$ degrees of freedom are available, meaning that we can make a random selection only $n-1$ times—the last number to be selected in order to make the sum total to τ is completely determined by the preceding $n-1$ random numbers.

But then: why $n-1$? Why not $n-2$, or some other number?

And also: how can we make sure that all $n-1$ selections are indeed random? One way in which this constraint could be violated is if the sum were to total τ before we get to $n-1$; say, on the $n-2$ selection. In that case, we would have not one but two values determined by preceding selections, at points $n-1$ and n. So in such a case we would in fact *have fewer than $n-1$ random numbers. But the goal is to ensure that we have maximum randomness, or alternatively, that we have minimal determinateness, which means we are required to ensure that all $n-1$ selections are random in fact. Only the last (n) selection is allowed to be determined by prior selections.*

Ed. note 0.0.6 *Why maximize randomness? I think this is the link to the concept of "normal distribution of error".*

The arithmetic mean of a set of *n* numbers can be thought of in terms of redistribution of total across *n* "sample points". The most obvious way to think of this, as suggested by the formulaic definition of the mean, is in terms of a constant function that assigns an equal portion of the total to each sample point.

This means we think of the mean in terms of a *counter-factual* situation. We interpret the original distribution across the sample points as a *fact*; it's what the world (i.e. the set) looks like when the total is distributed according to some (unknown) rule. Then we can ask what the world would look like if the total were distributed according to some other rule. The arithmetic mean tells us what things *would* look like *if* the total were distributed according to a rule that assigns the same amount to each sample point; in other words, if the rule of assignment were a constant function.

There are two constraints on the redistribution. One is the obvious requirement that the sum total after redistribution must equal the sum total before redistribution. That is the point after all; we want to know just how things would look if the *same* total quantity were differently distributed.

The other constraint is less obvious: it requires that the redistribution be according to a *simple* rule. The "rule" must of course be a function; but that is not enough: the function must be expressible by a single equation. This rules out step functions, which must be expressed by two or more equations.

The motivation for the second constraint is best illustrated by example. If all we want is redistribution according to rule, then we can define any number of rules that assign a fixed quantity to each sample point. This *seems* to be what we're doing with the arithmetic mean (my reasons for saying "seems" will become obvious in a moment). On this view, each sample point gets exactly μ units.[11] In mathematical terms, this means we are using the constant function $f(x) = \mu$ as our distribution rule. Under this kind of rule, the quantity assigned does not depend on the sample point; they all get the same allotment.

But there are many other possibilities. Indeed, we can effect a fixed redistibution of this sort by using functions—that is, polynomical equations—of any degree. Recall that the degree of a polynomial in x is defined as the largest exponent of x that appears in the defining equation. So the constant function $f(x) = a$ has degree 0; it can be interpreted as $f(x) = ax^0 + 0 = 1 \cdot a = a$. A function of degree 1 has the form of $f(x) = ax^1 + b = ax + b$ — an ordinary linear equation, where the independent variable (in this case, x) has an exponent of 1.[12] A function of degree 2 has the form $f(x) = ax^2 + bx + c$, and so forth. The symbols $a, b, c, \dots$ in such equations are called *parameters*; the idea is that such equations are *schematic* forms that determine a *family* of equations (functions), one for each assignment of fixed values to the parameters. In other words, $f(x) = ax + b$ does *not* define a specific function; rather it expresses a general pattern or abstraction to which an infinite number of specific functions conform. Taking $f(x) = ax + b$ as an example, by fixing $a = 3, b = 7$ we obtain (by substitution) the concrete equation $f(x) = 3x + 7$; fixing $a = 1, b = 0$ we obtain the identity function $f(x) = x$. And so forth.

[11] We will always use μ to denote the mean; where necessary for disambiguation we'll add a subscript, e.g. μ_k.

[12] Remember that the domain of our functions is the ordered set of integers $[1 \,..\, n]$. So the allotment assigned to any sample point would depend just on where it is in that ordered set.

Now back to our specific question: what would our set look like *if* we were to use a first degree distribution function based on the arithmetic mean? We are given the general form $f(x) = ax + b$; the task is to find values for the parameters a and b such that the resulting concrete equation defines a suitable redistribution function.[13]

In this case, each sample point would be allocated a different quantity, since the value of $f(x) = ax + b$ depends on the value of x. The graph of an equation of this form is a line. More specifically, it is a line whose slope is non-zero; this is another way of saying it is not a horizontal line—a horizontal line is the graph of a constant function.[14]

Now there are an infinite number of first degree equations we could use to redistribute our total. Which one or ones do we want, and why? Here we have two constraints. One is that the total after redistribution must equal the total before redistribution. The other is that the redistribution must range over the same number of sample points as in the original set. Technically this means we are placing a constraint on the *domain* of our function. Normally, function definitions of the form $f(x) = ax + b$ assume implicitly that the independent variable x ranges over the real numbers $\mathbb{R}$. Since we are interested in a discrete set of sample points, this means we need to place a *domain restriction* on our function. Symbolically we can express this using the domain restriction operator[15] $\lhd$ as follows: $S_n \lhd f(x) = ax + b$, where S_n denotes a "selection function" that selects n points from the domain of f. The whole equation means, informally, "the function whose domain is determined by S_n and whose range is determined by f", or alternatively "the function defined by $f(x) = ax + b$, with x constrained to range over the set determined by S_n".

TODO: address discrete v. continous at beginning

Now the key question becomes: what is the definition of S_n? And more to the point, what motivates it?

Ed. note 0.0.7 *NB: we could also use various functions for S_n. Why is a uniform linear distrib best?*

TODO: plot some examples. Use n points from the unit interval with $\mu = .5$. Animate the interpretation of μ as we move from f^0 to f^3.

Ed. note 0.0.8 *This is domain restriction by rule. So we actually have two rules. The domain restriction rule selects (that is, distributes) points from the domain, which determines the points for which we*

[13] Note that an implicit requirement is that $a \neq 0$, since $f(x) = 0x + b = b$ defines a constant function. (Or more accuratly, it defines the form of a family of such functions.)

[14] It also cannot be a vertical line, since the graph of a genuine function cannot be a vertical line. Another way of looking at it is to observe that the infinite family of equations (functions) described by $f(x) = ax + b$ corresponds to the infinite number of lines in the plane *except* the (equally infinite) set of horizontal and vertical lines.

[15] This symbol is defined by the Z notation.

will compute values; and the “redistribution rule” is the function that assigns values to the selected points.

We can express this conveniently by naming these two rules the “selection function” and the “redistribution function”. Both are deterministic; the former fixes the domain of the latter; the latter fixes the values assigned to the selected points.

Explain the difference between a function and its values, and random variable “observations”. When we use counterfactual functions to interpret the arithmetic mean, we’re moving from random variables to determinate functions. But this means changing our domain as well; we replace “taking” n samples with evaluating functions over infinite domains, which obviously include more than n points. The mean comes out as characterizing the function is some critical way—as the value of a constant function, the midpoint of a linear function, and so forth. In other words, such counterfactual thinking allows us to treat *random data* as if *it were “produced” by a function. Then our original taking of samples, which was “observing” the* value *of a random variable n times, corresponds to sampling the domain of a function. And to make this work, we move from n random samples to n points in the function domain distributed linearly. I.e. we select n points from the domain such that the distance between any two adjacent points is the same.*

Intuitively, we are interested in the line that runs from the minimum value of the original set to its maximum value. By the definition of the arithmetic mean, this interval from minimum to maximum is guaranteed to contain the mean. In fact, the arithmetic mean of the original set will correspond exactly to the midpoint of this line.

But this is statistics; we are not so interested in fixed values. We want to know not only how things look under a different distribution rule; we also want to know how things might look under distributions rules that have a *random* component.

Ed. note 0.0.9 *Having moved from the original (possibly) random distribution to counterfactual deterministic redistributions, we restore randomness by another reinterpretation. Instead of thinking of our redistribution function as fixing values for domain points, we treat it as fixing the parameterization of random variables—as setting “knobs” on randomization devices. Each point is construed as a randomization device.*

But this means we have moved back away from the function concept. We no longer compute the value of the redistribution function for each point in the restricted domain. Instead we construct n randomization devices as parameterized random variables, where "parameterization" sets the range over which the r.v.s may vary. Then we "observe" each r.v. and sum the results.

An obvious problem: how can we guarantee that our results will sum to the required total? This is where degrees of freedom and the concept of bias enter the picture. Degress of freedom always leave some points whose allocation is determined by the preceding allocations. This means that once the randomizers have generated their values, we can compute how much "stuff" remains to be distributed and assign that determinately to the remaining points. So this mix of "free" and "fixed" points amounts to the optimal solution to the task of maximizing randomness while ensuring that the overall constraints are satisfied.

A simple question: why bother with all this? Why do we want to allocate a total across some set of randomization devices? Possible answer: because this represents fitting the observed values to a probability distrib. The real question for which we want an answer is: how would things look if we were to use the distribution functions defined by probability theory as our redistribution function? And which of those distribution functions yield the "best" redistribution—the one that comes closest to matching the original data?

Ed. note 0.0.10 *Order of exposition/illustration: start by making the entire total avaliable to each device. Problem is that first point can use up the entire total; then all the remaining allocations are deteministically fixed. One possible remedy: add clauses to the redistrib rule to the effect that allocation to point n depends on previous allocations; this is disallowed (why?). Next try: allocate according to simple functions of degree n, as above. Difference is allocation as parameterization rather than fixing of values. First try: linear alloc evenly distributed across all points. This induces bias (why/how?). Solution: degrees of freedom. Distribute across n− degrees of freedom. This makes those points random, and the remaining deterministic. This is the optimal (only?) redistrib method that satisfies the basic requirements and maximizes randomization.*

Ed. note 0.0.11 *Need a good, simple, intuitive account of the technical concept of bias.*

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.

Part VII

A Statistical Beastiary

Specific statistics, distribs, tests, etc. - how they are used in practice.

Chi Squared

To find the expected multiplicities, form the disjoint union of the marginals.

Remark 37 Not sure “disjoint union” is the right term. The idea is that we take the observed marginals, say A and X, and form the set of all pairs (a,x) by multiplying. Effectively this means we take all observed pairs (a,_) and (_,x), ignore the second and first members, respectively, and then “cross” them. So for every (a,_) we get X copies of (a,x), one for each (_,x). The result has multiplicity of $A \times X$. I think this is disjoint union but I’m not sure.

Part VIII

Causality

Factor Analysis

http://psycnet.apa.org.proxy.uchicago.edu/index.cfm?fa=browsePB.chapters&pbid=10694

Part IX

Appendices

Appendices

Bibliography