2017

Data Science for Social Good Europe

Summer Fellowship

A critical introduction to statistics
Part I: Foundations

Momin M. Malik

v2.1, 15 August 2017

Any views expressed here are my own, and do not necessarily reflect those of DSSG.

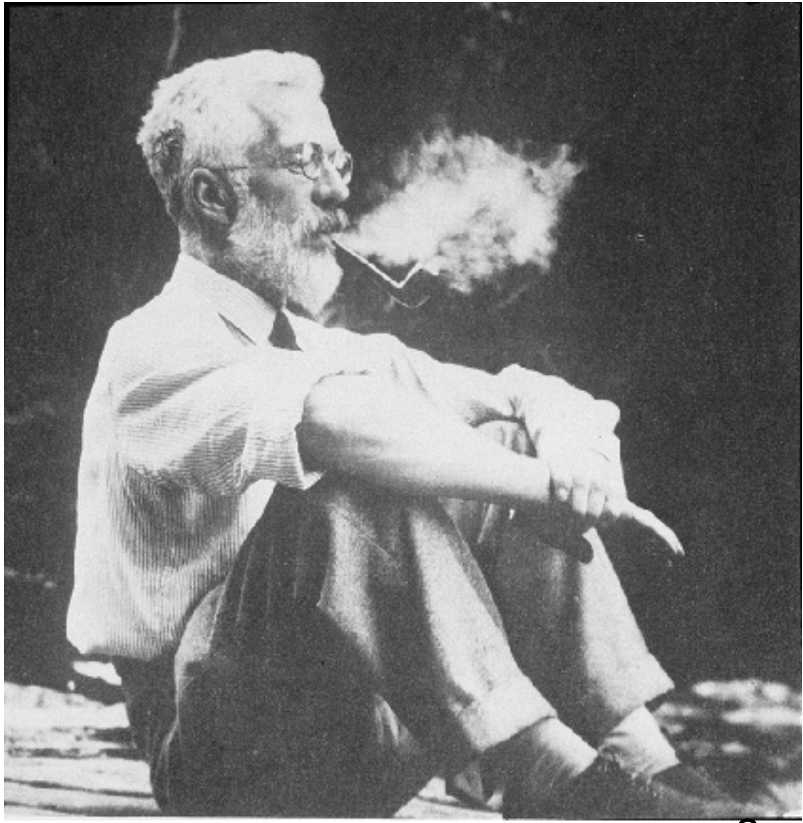
1. What is statistics?

What is statistics?

*“briefly, and in its most concrete form, the object of statistical methods is the **reduction of data**.”*

- R. A. Fisher, 1922, “On the mathematical foundations of theoretical statistics”

Fisher: Raging misogynist, paid shrill for the tobacco industry… and one of if not the greatest inventors of statistical practice.



What is statistics?

“A quantity of data, which usually by its ***mere bulk*** *is* ***incapable of entering the mind****, is to be replaced by* ***relatively few quantities*** *which shall* ***adequately represent the whole****, or… as much as possible… of the* ***relevant information*** *contained in the original data.”*

- R. A. Fisher, 1922, “On the mathematical foundations of theoretical statistics”

What is statistics?

A “statistic” (singular) is defined as *a function of the data*.

The discipline of Statistics is about *defining* “relevant information,” and finding functions to capture it.

How does it do so?

What is statistics?

I understand statistics as:

*The use of probability as a model for variability in the world.**

This talk is about this idea.

* Technically, *“Probability is used in two distinct, although interrelated, ways in statistics, phenomenologically to describe haphazard variability arising in the real world and epistemologically to represent uncertainty of knowledge.”* D. R. Cox, “Role of models in statistical analysis” (1990). For now I focus only on the former.

Probability as a model for variability

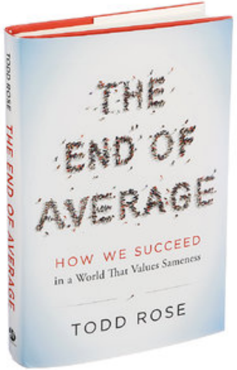
This implies a philosophical commitment:

There are distinct entities in the world that, despite being different, are similar in some way.

As a corollary, we can thus learn about one thing by studying other things (and eventually, make statements about not-yet-seen entities based on the study of seen entities).

Implications

Note that it is not necessarily intrinsically true, and that by holding this belief we may actually be imposing it on the world!



Averages are one way to summarize this similarity. For the problem with this, read the first chapter of Todd Rose’s *End of Average* (available from thestar.com as “When the U.S. Air Force discovered the flaw of averages”).

Or, to summarize, a true “average man” (Adolphe Quetelet’s *l'homme moyen*, 1835), who is average in all aspects, would be quite peculiar!

2. How does statistics use probability to model variability?

First: *the connection is not at all natural!*

"It is remarkable that a science which began with the consideration of ***games of chance*** *should have become the most important object of human knowledge."*

- Pierre-Simon Laplace, *Théorie Analytique des Probabilitiés* (1812)

And the connection seems quite distant

- In statistical practice, we almost never deal with probability statements, e.g., $P(X=0 \text{ AND } Y=2) = 0.1$
  - And even more rarely do we manipulate such statements
- Why, then, do courses in statistics always start with probabilities?

And the connection seems quite distant

- Because probability motivates everything in statistics even if we don’t see it in the final models
- E.g., regression is a “conditional expectation”
- So, unfortunately, it is necessary to cover even though it is a tiny part of practice.

Random variables

- Core concept is that of a *random variable*.
- What is a random variable?
- A die that has not yet been rolled. It represents *unrealized possibilities*.

What is the connection?

- Take users rating movies. *The rating that a user could (or will) give to a movie is conceptualized as a random variable*.
- Ratings that have already been made are *realized values* of that random variable.

What is the connection?

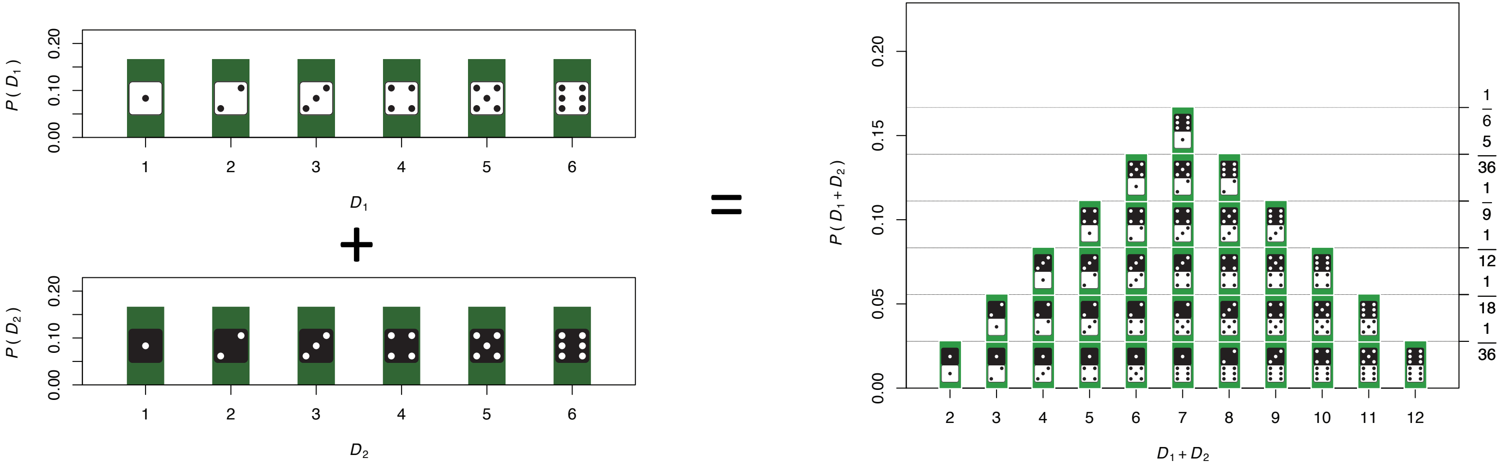
- Of course, different ratings are the result of *variability* in human tastes, not a “draw” from a “random variable.”
- So when we describe movie ratings as a random variable, we are using probability as a *model* for variability.
  - (Whether variability is the result merely of imperfect knowledge, or whether randomness *actually* exists, is something we can’t ever really know [Nasim Nicholas Taleb])
- This still doesn’t yet answer *how* random variables represent observed variability…

Back to random variables

- Write random variables as uppercase, e.g., X.
- Write a realization as lowercase, e.g., x.
- Probabilities are written as $P(X = x)$, the probability that the random variable will be realized as one of its particular outcomes. As shorthand, we (confusingly) sometimes write just $P(X)$, or $p(x)$.

Random variables → Probability distributions

- The remarkable thing about probability is that it provides the notion of randomness (the set of *possible outcomes*) having a “shape.” This is a probability distribution. Random variables are defined by their probability distribution
  - Note that random variables are one thing, the equation for their distribution are another. Annoyingly, no easy way to go back and forth.



Linking probability to the world

- But what does it mean for something to have a “probability”? Either the sum of a roll of two 6-sided dice comes out as 7 or not!
- Have to have some link between *possible outcomes* (which is abstract and mathematical) to what we actually observe in the world.

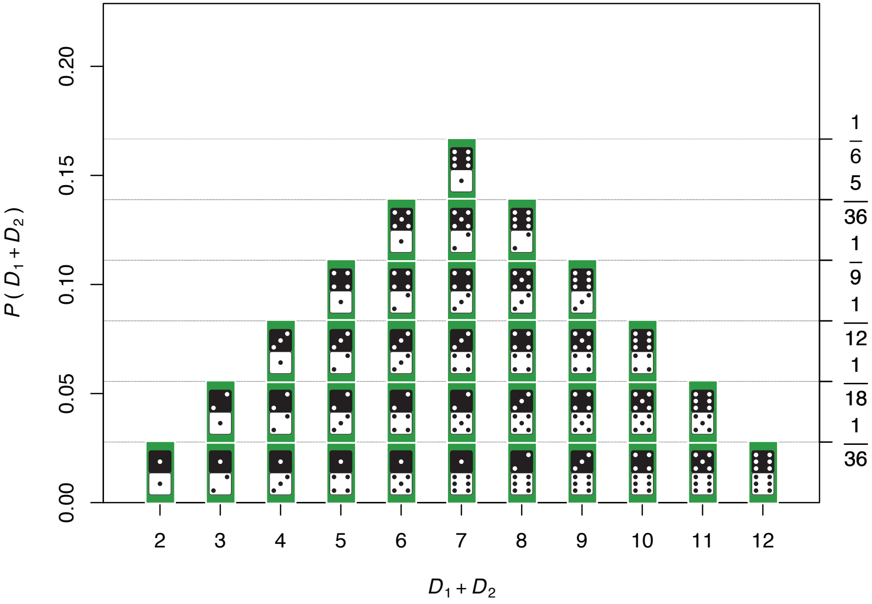
Linking probability to the world

- Here is one distinction between Frequentists and Bayesians:
  - Frequentists link probability to the real world in terms of *long-term frequency*: what happens over multiple runs.
  - We'll return to this later (what does long-term frequency mean for one-off events??)
  - Bayesians link probability to the real world via what we think, as "reasonable expectations" or "personal beliefs."

Discrete probability distributions

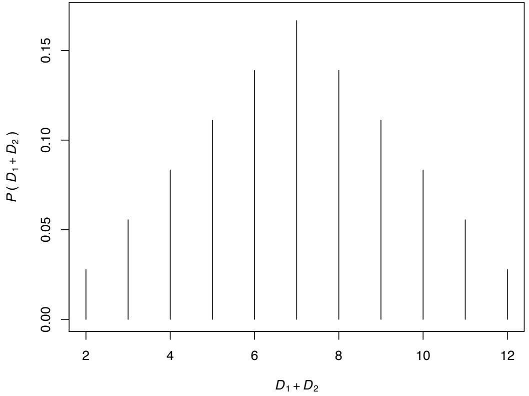
- A discrete set of outcomes (like dice) is comparatively easy to understand: the number of ways to reach each possible outcome
- $P(D_1+D_2 = s)$, where s is one of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. So, $P(D_1+D_2 = 2) = 1/36$, $P(D_1+D_2 = 3) = 1/18$…
- Can write as a vector, where each entry is the possible value:

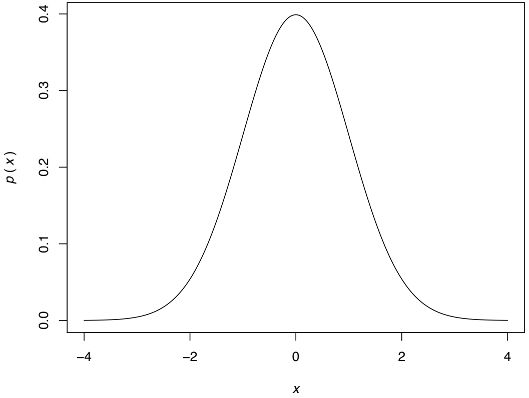
  $P(D_1+D_2 = (2,3,\ldots,12)) = (1/36, 1/18, 1/12, 1/9, 5/36, 1/6, 5/36, 1/9, 1/12, 1/18, 1/36)$



Continuous probability distributions

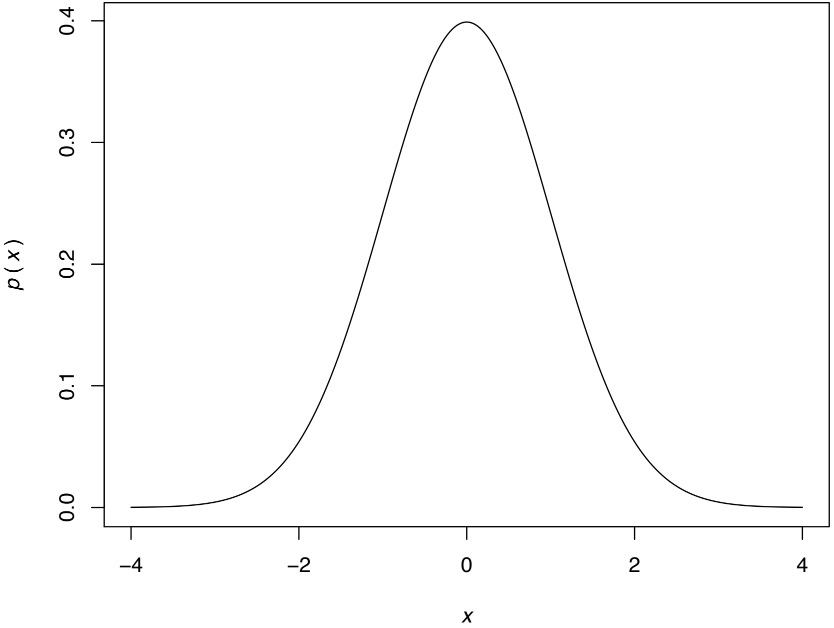
- But we also have *continuous probability distributions*, for random variables that can take any decimal value. These are strange.
- Unlike a 6-sided die, which has $P(X = x) = 1/6$ for $x = \{1, 2, 3, 4, 5, 6\}$, for a continuous probability distribution (a random variable that can take any decimal value), $P(X = x) = 0$ for all x.





Continuous probability distributions

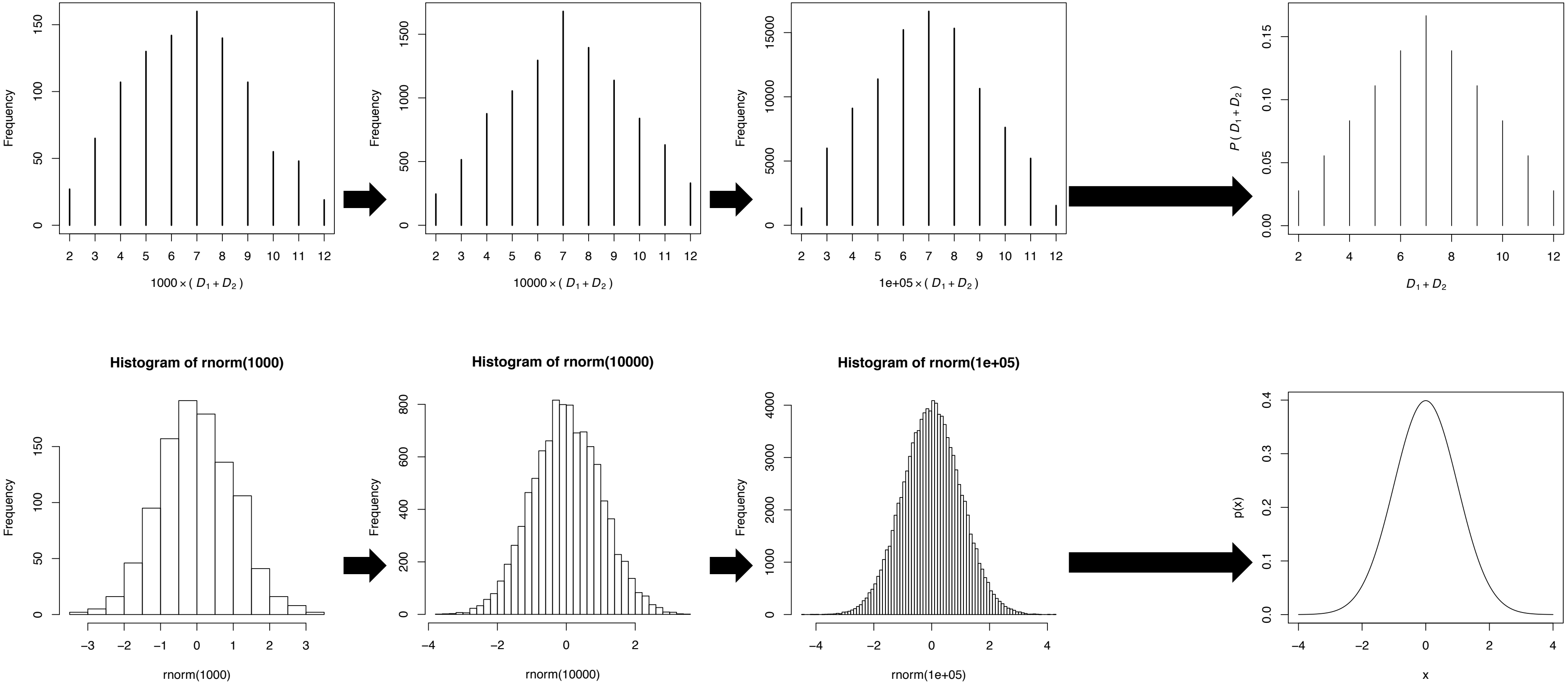
- The difficulty with understanding continuous probability distributions is the same as with infintesimals in general:
  - What does it mean for a substance to have a “density” of 1g/cm at a single location?
  - For an object to have a “velocity” of 80km/hr (in a direction) at one instant, frozen in time?
  - For a line to have a “curvature” at a single point?
- Infintesimals are counterintuitive, but we’ve learned to work with them
- *P*(*X* ≤ *x*) does make sense. But this is a function of *x*, and we can differentiate it to *p*(*x*) to get something equivalent to *P*(*X* = *x*).



Distributions are theoretical objects

- Both discrete and continuous distributions are hypothetical *underlying theoretical objects*. They do not exist in the real world; we appeal to them as explanations
- Frequentists have a notion of “asymptotics”: as we get more and more data from the same data-generating process, it will (at infinity) “converge” to the distribution
- For processes with discrete possible outcomes, we can just count instances. For processes with continuous possible outcomes, we have to bin observations and make histograms.

Distributions are theoretical objects



3. Statistical inference: From data to the data-generating process

Statistical inference: The big picture

- If you’ve taken a statistics class, you probably saw a picture like this (from Robert E. Kass, “Statistical inference: The big picture”, 2011):

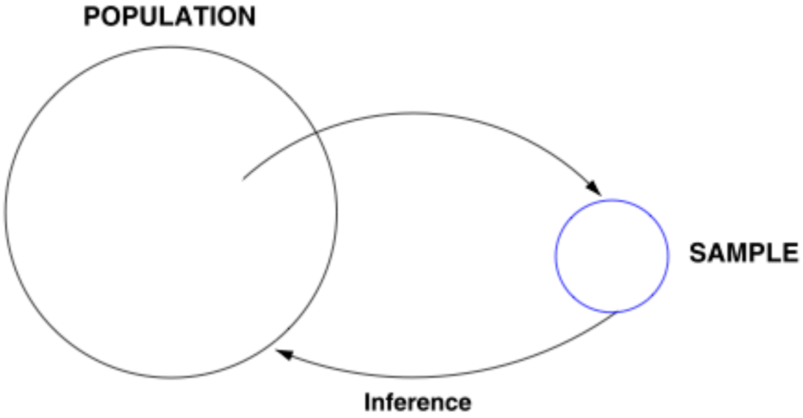


FIG. 3. *The big picture of statistical inference according to the standard conception. Here, a random sample is pictured as a sample from a finite population.*

Statistical inference: The big picture

- This is wrong. (Or at least, very limited).
- Again, quoting Fisher (1922):
  - *"This object [reduction of data] is accomplished by constructing a <u>hypothetical infinite population</u>, of which the actual data are regarded as constituting a random sample."*
- This is quite metaphysical! We don't want this in science.
- Jerzy Neyman and Egon Pearson came up with a way around this: there is no infinite population, only the larger "general" population, and *the randomness lies only in the process of sampling* (not in anything like realizations over multiple possible worlds). No metaphysics!

Statistical inference: The big picture

- But what if we have the entire population of interest? Then are we no longer doing statistical inference?
- No. A more honest, and general notion of statistical inference is that inferences are not to the general population, but to an underlying *data-generating process*. The randomness is not in sampling, but in the realization of a process that could have turned out differently.
- Unfortunately, this either requires a metaphysical commitment to multiple possible worlds.
- Or, for Bayesians, “reasonable expectation” (objective Bayes) or “personal belief” (subjective Bayes).

Statistical inference: The big picture

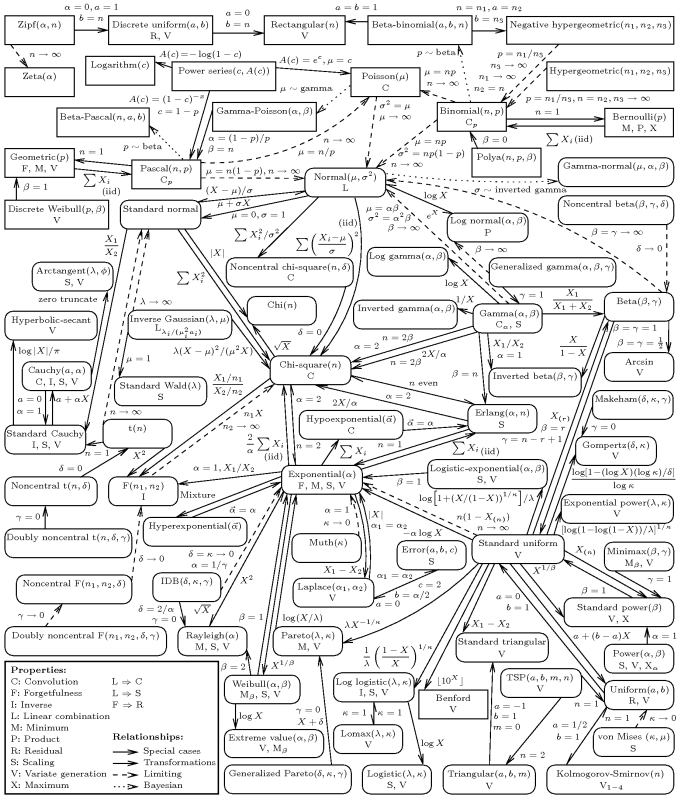
- The idea of statistical inference is that the “data-generating process” (random variables and interactions between them) is an *underlying theoretical object*.
- Using finite observations, we identify that object with some amount of certainty (where quantifying uncertainty also comes in).
- Once we have identified the object, we have understood the phenomenon.

Distributions zoo

- Certain probability distributions arise from certain *processes*.
  - Normal distribution: *continuous additive process*.
  - Binomial distribution: *binary additive process*.
  - Log-normal distribution: *multiplicative process.*
  - Poisson distribution: *counting process*.
  - Exponential distribution: *waiting times*.
  - Weibull distribution: *multiple failure process*.
  - Beta-binomial distribution: *Polya urn process* (don't ask.)

Distributions zoo

- One approach to statistics: come up with a distribution for each new processes.
- Usually nor worthwhile, since distributions are related (see diagram to right, Leemis & McQuestion, “Univariate distribution relationships”, 2008).
- Dangerous to infer the *process* from a distribution!
- Usually, we are more interested in the deterministic parts of data-generating processes, or relationships between random variables with simple distributions



Summary

- *Variability* is something we observe in the world
- *Probability* is a mathematical abstraction (that originated around gambling)
- Statistics describes variability in terms of probability.

4. Where does data come in?

The likelihood principle

- The *likelihood principle* is fundamental to statistical thinking.
- Take a the functional form of a probability distribution:

$$p(x) = 1/\sqrt{(2\pi\sigma^2)}\ \exp(-(x-\mu)^2/2\sigma^2)$$

- Take a set of data points, x_1 = -1.38, x_2 = -0.44, x_3 = 1.64, x_4 = -0.25, …
- *Invert things*. Instead of asking what is the probability of x_1 being -1.38, ask: what is the *most likely* value of μ that generated an x_1 of -1.38?
- Reconceive the functional form as $L(\mu)$, a function of μ
- Now called the *likelihood*.

Maximum likelihood

- By laws of probability, the probability that x_1 = -1.38 AND x_2 = -0.44 AND x_3 = 1.64 AND x_4 = -0.25 is the product of their individual probabilities.

$$p(x_1, x_2, x_3, x_4, \ldots) = p(x_1)\, p(x_2)\, p(x_3)\, p(x_4) \ldots$$
$$= \prod_i 1/\sqrt{(2\pi\sigma^2)}\ \ \exp(-(x_i - \mu)^2/2\sigma^2)$$

- Rewrite as a function of μ turns this into a likelihood:

$$L(\mu) = \prod_i 1/\sqrt{(2\pi\sigma^2)}\ \ \exp(-(x_i - \mu)^2/2\sigma^2)$$

- Use calculus to find the μ that will maximize the likelihood: turns out to be $n^{-1}\sum_i x_i$, the sample mean!
- Why do we care?
- The idea is that once we find μ, we have understood the process: *we cannot reduce things further than knowing the shape of the uncertainty*.

5. Why use statistics?

Reason 1: Deal with variability

- If we believe that there is variability in the world, and that data has value (although it is perfectly defensible, if not popular or acceptable, to challenge both of these!), then it follows that we have to use something like statistics to manage data.
- What are alternatives?

Reason 1: Deal with variability

- Descriptives (although these can involve low-level statistics, for doing summaries): limited
- Conceptual mathematical modeling: no data
- Simulation modeling: not models of data, don't give individual predictions
- Not accounting for variability: violates our belief
- Not doing mathematical modeling at all, but qualitative analysis: Should also do this, but can't make use of administrative and trace data
- Game theory, adversarial learning, etc., where variability is from decisions of an "opponent" (could be nature): in practice, can look very similar

Reason 2: It works

- Paul E. Meehl, *Clinical versus statistical prediction* (1954) managed to convince physicians that randomized control trials were superior to clinical judgment (which is not at all obvious)
- Even “improper” linear models work better than expert judgment
- (Note: this is not so simple of an argument. When modeling *doesn’t* work, as it frequently doesn’t, is it a failure of modeling, or the modeler? Saying the latter seems like apologetics. But also, if we can’t have modeling without modelers, does it matter?)
- (Also, if modeling works in places like finance, is it because it actually works, or because finance is a rigged game?)

That’s all for now!

- Next time(s):
  - Probability up through conditional expectation
  - Regression as conditional expectation
  - Nonparametrics
  - Misspecification: omitted variable bias, and dependencies
  - Prediction vs. explanation
  - Causality