Mathematics for Policy and Planning Science

Stephen Turnbull

Graduate School of Systems and Information

Lecture 2: April 23, 2018

Abstract

Introduction to set theory, probability, and statistics.

Conditional probabilities

- We define two events to be *independent* if $P[A \cap B] = P[A]P[B]$.
- We define the conditional probability of an event B given an event B as $P[B \mid A] = \frac{P[A \cap B]}{P[A]}$.
  - A and B are independent if and only if $P[B \mid A] = P[B]$ and $P[A \mid B] = P[A]$. This is a theorem, not the definition.
- *Bayes' Law* states $P[A \mid B] = \frac{P[B|A]P[A]}{P[B]}$. Bayes' Law is also a theorem, not an axiom.

Why You Should Understand Probability

- Ordinary probability is not very useful for calculations in business. Numerical assessments of probability are hard to get and inaccurate (if you ask several experts the spread is generally large).
- It's important to convince yourself that the probability laws make sense and are *the right way* to work with "likelihood" when you can.
- In fact, people *frequently* violate the laws of probability in their assessments. This fact is one of the foundations of *behavioral economics*, and has some impact on *psychometry* (heavily used in marketing and organizational behavior applications).

The "Linda Problem"

Quoted from the description in D. Kahneman, Thinking, Fast and Slow*, Ch. 15.*

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was very concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

Rank the following additional descriptions of Linda in order of probability:

1. Linda is a teacher in an elementary school.
2. Linda works in a bookstore and takes yoga classes.
3. Linda is active in the feminist movement.
4. Linda is a psychiatric social worker.
5. Linda is a bank teller.
6. Linda is an insurance salesperson.
7. Linda is a bank teller and active in the feminist movement.

The Taxicab Accident

Quoted from the description in D. Kahneman, Thinking, Fast and Slow*, Ch. 15.*

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- 85% of the cabs in the city are Green and 15% are Blue.
- A witness identified the cab as Blue.The court tested the reliability of the witness under the circumstances that existed the night of the accident and concluded the witness identified each one of the colors 80% of the time and failed 20% of the time.

What is the probability that the cab involved was Blue rather than Green?

Random variables

- It is often useful to identify the elements of the largest "interesting" partition of Ω as a set of *states*. If you know the probability of all states, you can compute the probability of all events made by combining them.
- The other thing you can do when Ω is treated as a set of states is define a *random variable* $X : \Omega \to S$ for some set S (typically numbers, vectors, or functions of time).
- We define the *distribution* (or *cumulative distribution function*) of a random variable $X : \Omega \to S$ to be the function $F(c) = P[\{\omega \mid X(\omega) \leq c\}]$ for $c \in S$. (Obviously S must be a set of numbers. It's possible to generalize, but we don't need to for this class.)
- We also define the *(probability) mass function* $f(c) = P[\{\omega \mid X(\omega) = c\}]$ or *probability density function (pdf)* $f(c) = F'(c)$ for $c \in S$ (depending on whether F is a step function or differentiable). *Note*: F always exists but f may not.

Understanding random variables

- A random variable allows us to express numerical uncertainty, such as when we wish to predict a stock price in the future.
- The primitive events can be anything; in fact in statistics we usually completely ignore them.
  - We can do that once we have defined the random variable's distribution.
- They are used so that we can understand concepts like independence and mutual exclusion for "random numbers."

Composing Random Variables

- If $f : S \to T$ is a function, then we can define a new random variable $f(X) : \Omega \to T$ by *composing* f with X.
  - For example, in a survey question we may ask a subject about some property of a product: choose the word that best expresses your feeling from "love," "like," "indifferent," "dislike," or "hate." Then the set S is the set of feelings {love, like, indifferent, dislike, hate}. We convert this from words to numbers between 1 and 5 using a 5-level *Likert scale*: love = 5, like = 4, indifferent = 3, dislike = 2, hate = 1. The set T might be defined as integers or real numbers, or it might be defined precisely as $T = \{1, 2, 3, 4, 5\}$. Each approach has advantages and disadvantages.
  - "Decomposing" the Likert scale into X and f has the advantage that it's easier to remember that quantitive measurement of feelings is difficult.

Independence of r.v.s

- In statistics, we often need several random variables whose probabilities are related to each other. In order to relate the probabilities of two r.v.s X_1 and X_2, they *must* have the same state space, but the target sets can be different: $X_1 : \Omega \to S_1$ and $X_2 : \Omega \to S_2$. Then we define the *random vector* $X = (X_1, X_2) : \Omega \to S_1 \times S_2$, such that $X(\omega) = (X_1(\omega), X_2(\omega))$ (same ω!)
- Define the *joint (cumulative) distribution* of two random variables X_1 and X_2 to be the function (of two arguments)
  $F(c_1, c_2) = P[\{\omega \mid X_1(\omega) \leq c_1 \text{ and } X_2(\omega) \leq c_2\}]$.
- Two r.v.s $X_1 : \Omega \to S_1$ and $X_2 : \Omega \to S_2$ are *independent* when
  $F(c_1, c_2) = F_1(c_1)F_2(c_2)$ for *all* $(c_1, c_2) \in S_1 \times S_2$.
  - **Warning:** In modeling, don't just couple two r.v.s without expanding the state space. They won't be independent! Rarely a problem in theory, this frequently *frequently* catches researchers creating simulations. Some of the things you believe can "really" happen are *impossible* in the simulation!
- Conditional distributions are defined in the same way using conditional probabilities. Again you need to be careful about state spaces.

Random variable example

- Consider a set Ω of primitive events, and a probability function for them. *E.g.*, a colored die with `red`, `orange`, `yellow`, `green`, `blue`, and `violet` sides, and the uniform probability

$$\begin{aligned} P(\texttt{red}) = P(\texttt{orange}) = P(\texttt{yellow}) &= \\ P(\texttt{green}) = P(\texttt{blue}) = P(\texttt{violet}) &= \frac{1}{6} \end{aligned}$$

- A random variable is a function $X : \Omega \to Z$ from the primitive events to some set, typically the real numbers R:

$$\begin{aligned} X(\texttt{red}) = 0, \quad & X(\texttt{orange}) = 1, \quad X(\texttt{yellow}) = 2 \\ X(\texttt{green}) = 0, \quad & X(\texttt{blue}) = 1, \quad X(\texttt{violet}) = 0 \end{aligned}$$

Related random variables

- We often define several random variables on the same primitive events, like $Y : \Omega \to R$, which is different from X:

$$\begin{array}{lll} Y(\texttt{red}) = 0, & Y(\texttt{orange}) = 0, & Y(\texttt{yellow}) = 0 \\ Y(\texttt{green}) = 0, & Y(\texttt{blue}) = 1, & Y(\texttt{violet}) = 0 \end{array}$$

  A random variable like Y that takes only the values 0 and 1 (and nothing in between) is called a *dummy variable*. Dummy variables are frequently used in statistics to analyze qualitative properties of sample observations.

- We can define one random variable from another: $Z = X^2$:

$$\begin{array}{lll} Z(\texttt{red}) = 0, & Z(\texttt{orange}) = 1, & Z(\texttt{yellow}) = 4 \\ Z(\texttt{green}) = 0, & Z(\texttt{blue}) = 1, & Z(\texttt{violet}) = 0 \end{array}$$

Simple Statistics

- Probability theory and statistics are quite different subjects in mathematics. (Many consider them disjoint!)
- Probability theory works with well-defined but abstract constructs, to discover how to do a calculation (broadly defined). We use whatever variables are convenient for the calculation.
- Statistics tries to infer some basic properties of a population or process from given data. The data may or may not directly convey the information we need. For that reason statisticians are typically very concerned with the "type" of data.
- Statistical data is summarized using *distributions*.
- Statistical distributions obey the same rules as probability distributions with one exception: in statistics, we often use *counts* rather than fractions, so the rule $P[A] \leq 1$ is relaxed.
  - A statistical distribution of counts is called *absolute*.
  - A statistical distribution of fractions is called *relative*.

Qualitative data and histograms

- In qualitative data, we have a set of values S, and that's all we know about it. For example, we could consider the nationalities of the members of this class: at least one American, several Japanese, several Chinese, and possibly others (which we should specify by name).
- The most basic operation on a qualitative data type is to partition it into subsets. In the nationalities example, I'd probably use "Japanese," "Chinese," and one subset "Other" including all the rest.
- The most basic operation on qualitative data is to construct a histogram, which is a function from the subsets in a partition to the frequency (for real data) or probability (for a probability model) of each subset. How many Chinese, Japanese, and Other.
  - Histograms are usually presented as bar charts or pie charts.
  - For purely qualitative data, the histogram is the closest approximation to a cumulative distribution possible.

Qualitative data and statistics

- The only statistics that can be computed are the *mode*, which is the *class* the occurs most frequently (not the frequency!), and the *Pareto distribution*, which is a histogram in which the classes are ordered from most frequent to least frequent. (Often "Other" is placed at the end anyway.)

Ordered data

- Data may be ordered in some natural way. *E.g.*, although we rarely think of colors as numbers, there is a natural order in terms of wavelength, or more poetically, the rainbow order. Similarly, strength of emotion can usually be judged in terms of more or less, but not assigned arithmetic values.
  - Many statisticians think of ordered data as "quantitative," but I find it more natural to consider it a separate kind.
- For ordered data, we define quantitative measure in terms of the distribution itself. For example, we often talk about the "bottom half" of the class, or the "top 10%" of the sales force. This practice is formalized as *fractile rank* (or *percentile rank* when expressed as percentages), where
  - the *r-fractile* is the value c such that $F(c) = r$, where F is a cumulative relative frequency distribution,
  - the *fractile rank* is r.

Statistics with ordered data

- Histograms and Pareto distributions may be used as with qualitative data.
- Special fractiles: median, quartile, decile, percentile.
- We often use ranges as measures of variability, such as the *interquartile range* to indicate "where most of the data is." (which isn't so far from ± 1 standard deviation for quantitive data), and the *interdecile range* (from the bottom 10% to the top 10%, useful in measuring the extremes of human inequality). But you should remember that the *difference* is meaningless for pure ordered data.
- With several ordered (but not cardinal) variables, the *Pearson rank correlation coefficient* can be calculated. This is an ordinary correlation coefficient, but it is calculated with the *rank* of each value rather than the actual value.

Cardinal data

- Often loosely called "quantitative" data, but most statisticians consider ordinal data to be quantitative.
- With cardinal measurements we can compute the usual *moments* (mean, variance, and so on, as well as the *non-central* moments where we do not transform by substracting the mean.
  - Variance is often transformed by square root to *standard deviation*.
- For two random variables we can compute *covariance*. We can also generalize to more than two by computing pairwise covariances and arranging them in a matrix, the *covariance matrix*.
- There is a huge variety of specialized statistics used with cardinal data, which you will meet in appropriate domains (*i.e.*, other courses).

Standardized data

- It is often useful to *standardize* cardinal data, by subtracting the mean and then dividing by the standard deviation.
  - Also called *normalization* (frequently) or *canonicalization* (incorrect and rare).
- Sociology-oriented packages like SPSS will do this automatically for qualitative and ordered data in regressions, and report coefficients as β (usually the standardized data) and B. Economics-oriented packages like E-Views usually report only the coefficients for unstandardized data.
- **N.B.** In nonlinear models these *different statistical models* because they change the relation between variability of independent variables and the error term!

Nonlinear/nonsmooth statistics

- The statistics we most frequently use in business and economic applications are linear and smooth (like the mean). This is very convenient for the mathematical theory: we can use linear algebra and calculus.
- Typical applications are linear regression (which can be derived in several ways: as the *best linear unbiased estimator* of the coefficient vector in a linear model, as the *least squares estimator*, or as the *maximum likelihood estimator* in a linear model with normal disturbances.
- These theories were originally developed in the context of linear models and the normal distribution. They have been extensively generalized, but there are several reasons why they may not be appropriate in many research cases.

Man is the measure of all things

- Old proverb
- Economics: elasticity
- Statistics: standard deviation
  - Standardization
  - Correlation coefficient: Motivation is somewhat similar to elasticity, but the "standard unit" is not the *level* of a variable. Rather it is the amount of variability.

Statistics, data mining, and big data

- For most Shako students, calculus and linear algebra are used in class as an aid to understanding theory rigorously (including the theory of statistics), but not used actively as research tools.
- Statistical tools (including data mining and big data) are used by almost all Shako students
  - Even those working in theoretical research, in example applications or to motivate their theories.

The "new statistics"

- The subject of statistics has changed dramatically in the last two decades.
  - Partly due to development of statistical theory.
  - Partly due to availability of new kinds of data (especially "big data" from sensor systems, POS (point of sale) data, and social networks).
  - Partly due to the diffusion of cheap powerful computers such as GPUs.

Choosing advisors

- Of course the Shako faculty who teach statistics are at the forefront of the new methodologies, but the applied faculty lag.
  - Keep this recent historical development in mind when you choose your AG.
  - Your principal advisor should be chosen for domain knowledge, but if you will do empirical work, I recommend you choose one advisor for their knowledge of the "new statistics".
  - You should also know that statisticians are increasingly specialized. At the least you should be careful to find out whether your proposed advisors research econometrics (best for relatively "hard" data such as prices and quantities), correlation analysis (for "softer" data based on subjective reports), or "big data" and machine learning algorithms.