Mathematics for Policy and Planning Science

Stephen Turnbull

Graduate School of Systems and Information

Lecture 2: May 11, 2020

Abstract

Introduction to set theory, probability, and statistics.

Introduction to Probability

- *Probability* is a numerical assessment of "likelihood" of *events*.
  - It is a mathematical model.
  - It obeys some convenient mathematical laws.
- It is useful in understanding "rational behavior":
  - If two actions A and B both result in the event "1000 yen profit," but with action A the result seems more likely (has *higher probability*) than with action B, it is "rational" to choose action A if only one of the two actions is possible.
  - In fact, this is an *axiom* of Laurence Savage's *theory of expected utility*.
  - In this sense, the economic theory of rationality is somewhat circular: we *define* rationality by the axioms, which we claim are "obviously" rational (in some undefined way!)
  - Such circularity is true of any set of axioms (such as the Peano axioms for the natural numbers).

Uses of Probability in Social and Economic Planning

- Risk analysis.
- Foundation of modern finance theory (especially of derivatives).
- Economics of insurance and investment (similar to finance theory, but adapted to the needs of the field).
- Most important, the foundation of statistical *inference*, and so the basis for most empirical work.

Assessing Probability I: Historical Frequency

- Where do the numbers come from?
- We can *measure* probability of an event of some type empirically by
  - counting the frequency with which it is observed in a repeated context,
  - comparing it with the frequency of occurance of the context, and
  - *assuming* that this relative frequency is a law that holds whenever the situation arises.
- An example of a "situation" is tossing a coin, and the event is that the coin falls "head up". It's easy to imagine repeating this situation many times to determine if the coin is "fair."

Assessing Probability II: Symmetry

- We can derive probability theoretically from a set of *symmetric* events by giving equal probability to each member of the set.
- A coin is "fair" if it obeys this symmetric law of probability, and therefore the probabilities of heads and tails are each equal to 1/2.
- The number of dots on the top face of a rolled die when it comes to rest are symmetric, with probability 1/6.
- The *sum of number dots* on two dice is not symmetric, but the *pairs of number of dots* is (if considered correctly, *i.e.*, as *ordered* pairs so that $(1, 2) \neq (2, 1)$).

Assessing Probability III: Subjectivity

- If we observe a rational actor who chooses to "bet" on event A rather than event B although we know that they are equally *profitable*, we infer that the actor assesses higher *probability* for A than for B.
- We say that such an agent has *subjective probabilities*.
- The agent may have reasons (such as an empirical frequency or a theory based on a symmetry) for the assessment but we don't know them. We often aren't sure the agent has reasons at all.
- If the agent tells us he doesn't care whether to bet on heads or tails with equal payoffs, we know she gives equal probability to each outcome.
- Individual humans are not very good at making such assessments. They not only have *high variance*, they also tend to be *significantly biased*. (Kahneman's *Thinking Fast and Slow* goes over the evidence for bias.)
- Remarkably, averaging assessments over many individuals reduces both *bias* (somewhat) and *variance* (dramatically).

Assessments and Mathematics

- In probability theory, we don't care where the assessments come from, we just calculate according to some rules.
- When applying probability to decision theory or to statistical study, we need to make some assumptions about existence and stability of the probabilities across trials.
  - These assumptions are derived from some domain of knowledge. As long as the assumption obey the laws of probability, probability theory can't help choose good assumptions. It can only rule out impossible (inconsistent) assumptions.
- This domain-specific process is called the *data generating process*.

Rules of Probability

- The rules of probability are based on *set functions*.
- Events are modeled as sets, which allows definition of rules for combining events logically.
- A *probabilistic model* is a function which gives a numerical value to certain sets, and obeys certain rules for the value of combined sets.

Example: Choosing a Chair

- A committee has 6 members: Ann, Bob, Carol, Don, Ed, and Frank.
- Nobody really wants to be the chairperson, nobody really hates the idea, and they decide to choose by lottery.
- Ann has a 6-sided die, and proposes the rule that each person has a number: Ann $\leftrightarrow$ 1, Bob $\leftrightarrow$ 2, Carol $\leftrightarrow$ 3, Don $\leftrightarrow$ 4, Ed $\leftrightarrow$ 5, and Frank $\leftrightarrow$ 6.

  Rolling the die will generate a number, and the corresponding person becomes chair of the committee.
- We intuitively ("naturally") think that each person can become chairman "1/6 of the time". Although we don't have to choose this as the "probability," it's convenient and symmetric, and reasonable given the symmetry of the die.
- "Ann becoming chair" is the way we describe a certain event in words. As a *set*, we describe it as "{Ann}" (not as "Ann"!) Then another event of some interest is "$\{\text{Ann}, \text{Carol}\}$," which we describe in words as "the chair is female." There may be many ways to describe the same event, such as "the chair wears high-heeled shoes".

Introduction to Probability

- Basic probability theory is based on numerical assessments of "likelihood" of *events*.
- Events are treated using *set theory*, and they can be combined and analyzed as sets are.
- Sets can be described explicit lists of elements (one by one), or as collections that satisfy properties. Element lists aren't very useful in probability theory.
- Sets described as satisfying a property connect events to logical propositions.
- We use the rules of logic to combine events, defined by such *properties* of the elements of the sets we think of as events.
- It is useful to consider all possible intersections of events, which give events with "few" elements, or satisfying very precise descriptions. A *partition* is a set of non-empty events such that the union of all of them is the certain event, and no two events in the partition intersect.

Assessing Probability

- Probability obeys some convenient mathematical laws.
- These laws determine the relationship of probability of an event to certain other events, but not the numerical value. There are three general approaches to assessing numerical probabilities:

  Frequency approach The relative proportion of times a specific event occurred in the past is taken as the probability. *Example: weather prediction.*

  Symmetry approach two events which are physically the same are assigned the same probability. *Example: rolling a die.*

  Subjective approach Let an expert estimate it, or use *revealed preference*: if an agent can bet on two events A and B, they have the payoff to the agent, and the agent chooses A, then A has higher *subjective probability* for the agent.

Operations on Sets

$A \cup B$	union	the set of things that are members of at least one of A or B; if both, count only once
$A \cap B$	intersection	the set of things that are members of both A and B
$A \setminus B$	set difference	the set of things that are members of A but not B
$A \subset B$	subset	every element of A is an element of B

Examples, where:

$$\begin{aligned} A &= \{\text{scissors, paper, stone}\} \\ B &= \{\text{needle, thread, scissors}\} \end{aligned}$$

$$\begin{aligned} A \cup B &= \{\text{scissors, paper, stone, needle, thread}\} \\ A \cap B &= \{\text{scissors}\} \\ A \setminus B &= \{\text{paper, stone}\} \end{aligned}$$

Set Operations on Events and Special Events

$A \cup B$	disjunction	the event where the event described by A, the event described by B, or both happens
$A \cap B$	conjunction	the event where both the event described by A, and the event described by B happen
$A \setminus B$	exclusion	the event where A occurs but B does not
$A \subset B$	implication	if A happens, then B does too
$\emptyset$	null event	"nothing happens", the *impossible* event, or *empty* event (here, "nothing" is *impossible* because "everything stays the same" is considered something happening)
Ω	certain event	the disjunction of all possible events, thus "something happens", also *sure* event

Probability and Special Events

- Each event has a probability (number) assigned to it. (In mathematics this is called a *set function*.) The probability of an event E is denoted $P[E]$. Some times "P" is spelled "Pr" or "Prob".
- The set of all elements of all events we are interested in is called the *certain event* or the *sure event*, and often denoted by Ω. The probability of Ω is 1.
- The empty set is called the *impossible event* and has probability of 0 (if you add no numbers, you have nothing, *i.e.*, zero). It is denoted by $\{\}$ or $\emptyset$.

Events and Probabilities

- Probabilities are not determined by mathematical theorems: they must be measured, deduced from assumptions, or estimated subjectively. However they do satisfy certain laws:
  - $P[\emptyset] = 0$ and $P[\Omega] = 1$.
  - $0 \leq P[A] \leq 1$ for all events A.
  - If $A \subset B$, then $P[A] \leq P[B]$ for all events A and B.
  - If $A \cap B = \emptyset$, then $P[A \cup B] = P[A] + P[B]$ for all events A and B.
- These are a sufficient set of axioms for countable Ω. There are many other probability laws that can be deduced from these.
- Note that these laws do not refer to elements!

The Sure Event

- It is very common in probability modeling to use uniform probability on some set Ω to structure the probabilities (as we did with the "chair selection problem").
- Usually these are finite sets, the unit interval $[0, 1]$, or products of such sets.
  - *Example 1*: Consider the chair selection problem and define Ω as the set of possible pairs of numbers from a red die and a blue die with equal probability for each pair of numbers. The event "Ann becomes chair" is defined by the set of pairs where the sum of the dice is divisible by 6, "Bob becomes chair" by all pairs among those remaining divisible by 5, and so on for Carol, Don, Ed, and Frank.
  - *Example 2*: Any increasing function $f : [0, 1] \to [0, 1]$ with $f(0) = 0$ and $f(1) = 1$ defines probabilities for all useful subsets of $[0, 1]$ via the Stieltjes integral (we will call it "cumulative distribution function"). It defines probabilities of intervals by subtracting the values at the endpoints, and (almost) everything else (of interest) by unions and intersections.

Homework 2, due May 18, 12:00 noon

Read and understand the following instructions on submission of homework. **If you do not follow them, you will not receive credit.**

Submit this assignment by *email*. Give the mail the subject "01CN101 Homework #<number> by <your name>" in *hankaku romaji* and send it to `turnbull@sk.tsukuba.ac.jp`. (This subject is necessary for automatically sorting incoming mail.) It should look like this:

```
Subject: 01CN101 Homework #1 by Stephen Turnbull
```

for Homework #1.

Make sure that the body of the email contains your *name* and *student ID number*.

If you are late, submit the assignment for partial credit. The later, the less credit you will receive. If you believe that the late submission is in part due to lack of care by the instructor, or some event (such as hospitalization) required your full attention for two full days or more, you may explain for additional credit. Otherwise, I don't care why it was late.

Please answer questions in plain text in the email. Exception: You may attach a

spreadsheet for the tables, including those used to construct distributions as well as the distributions themselves.

In class we showed that the distribution of sums of dots on a pair of dice is:

sum	2	3	4	5	6	7	8	9	10	11	12
frequency	1	2	3	4	5	6	5	4	3	2	1

Table 1: Sum of two dice

The distribution of ordered pairs is said to be *uniform*. For each pair, the frequency seen in a series of throws of the dice should be about the same. Theoretically, this occurs because each (ordered) pair occurs once in a list of pairs (not shown here – usually displayed as a square table with six rows and six columns). The frequency of sums is *nonuniform* (different frequencies for different values). Theoretically this difference occurs because there are often multiple pairs of dice that result in the same sum and the number of appropriate pairs varies according to the sum specified.

Homework 2(ii): Problems 1-2

1. Count the number of combinations of the number of dots on a pair of dice that achieves a particular:
   (a) product
   (b) difference
   (c) quotient

   of the pair and make a frequency table for each arithmetic operation (times, minus, and divide).

2. Did you notice any interesting similarities or differences between the frequency distributions? Did you notice any interesting similarities or differences between the random variables (*i.e.*, the function from the pairs of dice to the mathematical result)?

Homework 2(iii): Problems 3-4

3. Describe a model of "what is the gender of the first person to arrive in the classroom of 'Mathematics for Policy and Planning Science'" with an underlying set whose probabilities are *uniform*. What "model" means here is "What things do you count to determine the probability that the first person to arrive is female?"

   Do you think this model actually describes the probability that the first person to arrive is female accurately? If so, why? If not, why not?

4. Think of a practical or daily life application where you can explain observed rates of occurance with a model with an underlying uniform distribution, but a nonuniform distribution for the observed (or practically relevant) outcomes. What is the underlying set composed of? Why should its elements be uniformly distributed? What is the function relating the underlying things to observed outcomes?

Homework 2(iv): Problem 5

5. **(Optional)** Here's a brain-breaker for those of you who think you're good at math. The numbers on dice are all positive, and therefore it makes sense to take logarithms. There's a one-to-one relationship between numbers and their logarithms, so given a number we can find its logarithm, and it is the unique number with that logarithm. And given a number that is a logarithm there's a unique number it is the logarithm for.

   Now, after taking logarithms, multiplication becomes addition. Therefore we might expect that there should be a one-to-one relationship between the distribution of *sums* of two dice and the distribution of *products* of two dice. That is wrong.

   Explain why.

Homework 3, due May 18, 12:00 noon

Read and understand the following instructions on submission of homework. **If you do not follow them, you will not receive credit.**

Submit this assignment by *email*. Give the mail the subject "01CN101 Homework #<number> by <your name>" in *hankaku romaji* and send it to `turnbull@sk.tsukuba.ac.jp`. (This subject is necessary for automatically sorting incoming mail.) It should look like this:

```
Subject: 01CN101 Homework #1 by Stephen Turnbull
```

for Homework #1.

Make sure that the body of the email contains your *name* and *student ID number*.

If you are late, submit the assignment for partial credit. The later, the less credit you will receive. If you believe that the late submission is in part due to lack of care by the instructor, or some event (such as hospitalization) required your full attention for two full days or more, you may explain for additional credit. Otherwise, I don't care why it was late.

Consider the "chair selection problem" of *Example 1.*

1. What are the probabilities of each pair of *numbers*?
2. What events (as sets of number pairs) correspond to the events that each committee member is selected as chair?
3. Is it possible for all members to become chair? If not, which members are possible, and which impossible?
4. Explain why the definition says "all pairs remaining."
5. What are the probabilities of each member becoming chair?

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 correctly identified each 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: $Z = X^2$:

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

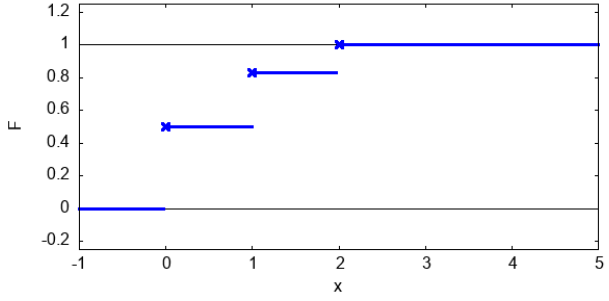
Distributions

For our random variables X, Y, and Z, it's convenient to use the *probability mass functions* to characterize the distributions.

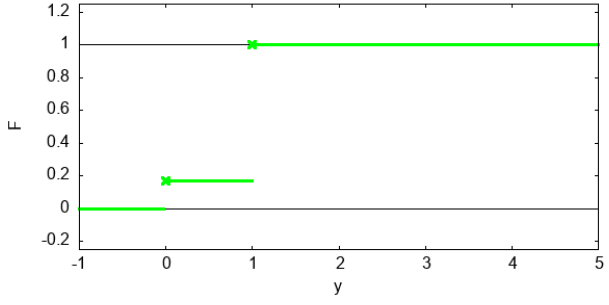
- X: $p(0) = \frac{1}{2}$, $p(1) = \frac{1}{3}$, $p(2) = \frac{1}{6}$, and $p(x) = 0$ for all other numbers x.
- Y: $p(0) = \frac{5}{6}$, $p(1) = \frac{1}{6}$, and $p(x) = 0$ for all other numbers x.
- Z: $p(0) = \frac{1}{2}$, $p(1) = \frac{1}{3}$, $p(4) = \frac{1}{6}$, and $p(x) = 0$ for all other numbers x.

Cumulative Distribution Functions

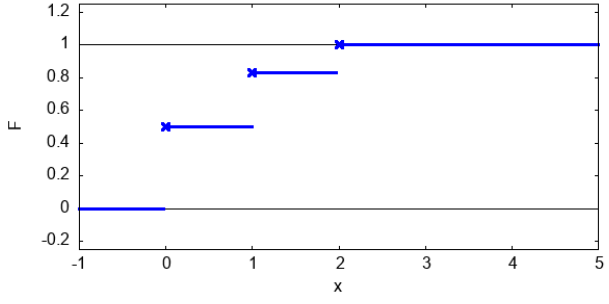
Cumulative distribution of the random variable X



Cumulative distribution of the random variable Y



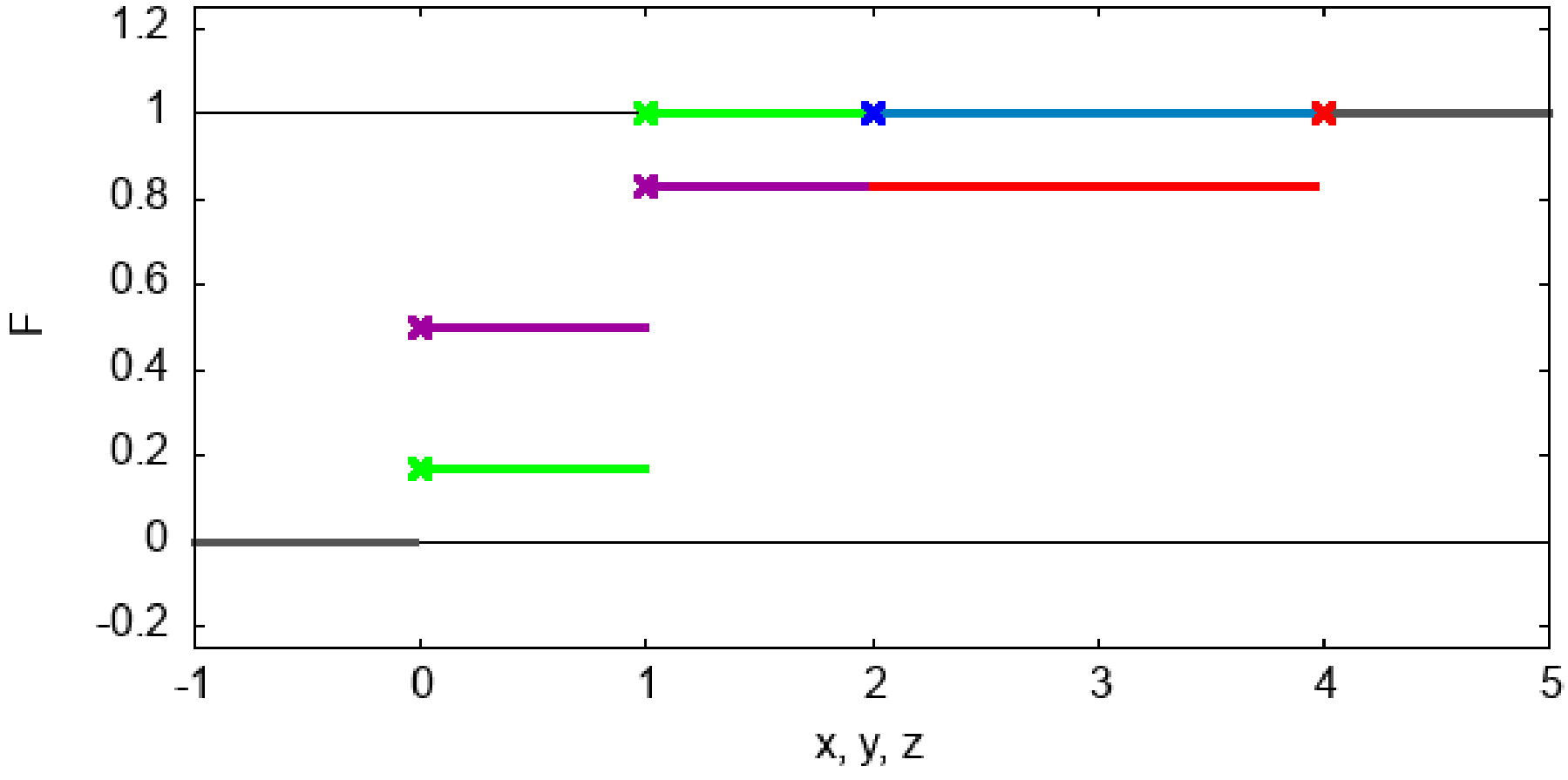
Cumulative distribution of the random variable Z



Combined Cumulative Distribution Functions

Cumulative distributions of the random variables X, Y, Z

The graph colors mix with each other, and may be hard to read



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, especially Bayesian statistics.
  - 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.

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). Remember that the *difference* is meaningless for pure ordered data, you need both *ends* of the range.
- 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.

Qualitative data: Is it ordered?

- In the nationality example, there's pretty clearly no "standard" order. Population, GDP, and life expectancy all give different orders!
- In data collected by design (*e.g.*, questionnaires), we can often **create** order by choosing words well (*Likert scales*).
- But how about Twitter text data: in a thread devoted to restaurants, what is the order of "tasty," "oishii," "umai," and "yabai"? Are you even sure "yabai" refers to taste?
- How about "yabai!" vs. "yabakunai?" in anything?!
- Recently in Israel, there was an election. The hard-right Likud candidate Netanahyu got the most votes. Generally we can talk about a left-right order in politics, but Netanahyu's coalition blew up because two even-farther-right parties got in a disagreement over military service.

Cardinal data

- Often loosely called "quantitative" data, but most statisticians consider ordinal data to be quantitative. I prefer to make the distinction explicit.
- 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 subtracting 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!

Ordinal data: Is it cardinal?

- We've already mentioned the ambiguous status of *Likert scales*.
- But "obviously" quantitive data like prices and quantities in economic data are not necessarily cardinal in the usual sense.
- Even economists use *elasticity* of demand and supply curves rather than *slope*.
- This can be handled with *data transformation* (in the case of elasticity, take the logarithm of both price and quantity data) or with *nonlinear regression*.
  - With economic growth and inflation, we probably don't want to use nonlinear regression.
  - With satellite photography, we probably don't want to transform data because it's 1000km from a reference location.

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. (You learn about this as you choose your research field. If a paper in your field uses linear models, *ask why*!)