NYU Department of Politics - Quant I
Fall 2013 - Prof. Patrick Egan

1 Lecture 1

- Housekeeping.
  - Syllabus:
    - Deliverables.
      - Attendance.
      - Class and lab.
      - Homeworks.
      - Exams.
    - Grades.
    - Help.
      - Introduce Andrew.
      - Propose Andrew's office hours; check to see if works.
    - Books.
    - Stata.
    - NYU Classes. [check.]
    - Hand out questionnaires.
- Why are you here? (i.e., why are you getting a Ph.D. in political science?)
  - You are here because you enjoy asking and answering questions about politics.
    - (If not, you're in the wrong line of work.)
  - So what kinds of questions do you find interesting? List on board.
- This is a course about quantitative analysis in political science.

  - Quantiative analysis takes its place alongside other methodologies in empirical political science. It is particularly helpful and appropriate for particular kinds of social scientific tasks.
    * Tends to be based on: *numerical* measurements of specific aspects of social and political phenomena; [contrast to: non-numerical]
    * Is more interested in developing and testing generalizable theories about *multiple* phenomena; [particular cases]
    * Seeks measurements and analyses that are easily *replicable* by other researchers. [difficult to replicate in its entirety.]
  - Contrast this to *qualitative* research [on all three aspects].
  - In addition, appropriateness of *qualitative* research depends on...
    * Hypothesis testing vs. hypothesis generation
    * Agreed-upon measures of concepts vs. those still up for debate
    * Analyst's willingness to apply less or more *structure* to the study of the phenomenon

- When political scientists work with data, we generally do so for three reasons:
  - What can we say about the data we have?
  - What can we say about the data we don't have (but we know is out there in the world) based on the data we do have?
  - What can we say about the data we'd expect to see under a hypothetical scenario, based on the data we do have?

- These activities generally require three kinds of statistics:
  - Descriptive
  - Inferential (from samples to populations)
  - Prediction (from models to hypothetical scenarios)
    * And lingering behind all of this is a kind of statistic we'll use all the time: the *test* statistic.

- What's a *statistic*?
  - It's a number that summarizes data.

1.1 Variables 101

-
  - We study **units**. They are the level at which we wish to make statements about a social process. Units are often also known as **cases**.
    - People, Counties, Nations, Dyads.
  - Units have **attributes**. An attribute is any characteristic of a unit that in theory might distinguish it from other units.
  - **Variables** are *logical* groupings of *mutually exclusive* attributes. The analyst assigns each attribute a **value**. We then say that a variable takes on a value for any particular unit.
    - The variable "hair color" takes on the value "red" for the unit Me.
    - The variable "party affiliation" takes on the value "Republican" for the unit Mitt Romney.
  - For ease of data manipulation in quantitative analysis, we typically assign **scores** to each potential value a variable can take on. The scores are simply numbers associated with each value. Sometimes the scores are meaningful in their own right; often they are not. But it's generally a lot easier to enter and manipulate data via numbers than via the words we use to describe values.
  - Note that all of these–which units, which attributes, how to group into variables, how to score the variables–are *choices* that must be considered and justified by the analyst. Typically there are conventions within subfields of political science that either must be adhered to (which is what we usually do), or if we depart from them we need to justify this departure. Sometimes the departure itself is a noteworthy innovation. Going to leave aside here a whole field of theory on how we move from concepts to **measures**. In this class, in most cases, we'll take the measures as given.

1.2 Levels of measurement

- Variables can be measured at various levels. We'll consider four such levels in our class:

  - Nominal
  - Ordinal
  - Interval
  - Ratio

- Variables measured at the **nominal** level take on values that cannot be ordered in any logical way.
  - Egs.

- Variables measured at the **ordinal** level taken on values that can be rank ordered, but that's it. We know nothing about how much more or less one value is than another.
  - Egs.

- Variables measured at the **interval** level take on values whose differences can be meaningfully compared. I.e., the interval between two values is meaningful.
  - Eg: Fahrenheit. The Year.

- Variables measured at the **ratio** level take on values such that the value of zero is meaningful in a specific sense: it means *nothing* of the quantity being measured..
  - Eg. income. height. age...number of wars...miles.

- Mathematical operations:
  - Nominal: equality.
  - Ordinal. equality. greater than or less than.
  - Interval: addition and subtraction; averages
  - Ratio: multiplication and division (i.e. ratios) - "twice as tall," "six times as wealthy"
  - Note that we generally have more information as we move up the ladder. Your first instinct should be to use measures that retain as much information about your units as possible.

- Three trickier cases:
  - Richter (ratio, although slightly tricky: $y_{Richter} = \log_{10}[\text{shaking amplitude}]$, so each unit represents a *doubling* of amplitude.)
  - Celsius has a meaningful zero (the freezing point of water), but it is *not* ratio level. Why? Because "zero degrees Celsius" $\neq$ zero of the quantity "temperature." [E.g., is a 30-degree Celsius day (86F) thirty times as hot as a one-degree Celsius day (32F)? Not meaningful.] Contrast to "zero miles" or "zero years old." Celsius is interval level. Contrast to Kelvin, measured at the absolute zero. Zero degrees K = -273.15 C; 1 degree K = 0 C, and so on.
  - Latitiude, longitude: for all intents and purposes, ratio. Of course, zero on the latitiude scale doesn't mean "zero distance." It means we are at the Equator. But as a unit of measure, to travel two degrees latitude is to travel twice as far as one degree latitude. If you travel from the Equator to Vermont, at 45 degrees latitude North, you've covered *half the distance* to the North Pole (at 90 degrees latitude). Therefore ratio.
- A special case: the dichotomous variable. Two ways to consider it: as a nominal variable with two categories, or as a ratio variable with two values, zero and one. E.g. "gender" versus "female."
- Choice of level of measurement is often up to the analyst. Many underlying concepts yield several choices for levels of measurement. Again, your first instinct should be to use measures that retain as much information about your units as possible. E.g.:
  - Location of residence: region of country [nominal], county [nominal], zip code [nominal!], latitude and longitude [ratio].
  - Hair color: common usage [nominal], amount of pheomelanin and eumelanin (hair pigments) [ratio]
  - Income: poor, middle class, rich [ordinal]; dollars per year [ratio]

1.3 Data structures

- The typical way that we store data (and really, the way that we think about data structures) is via a **data table**. Excel, any statistical software program. Each unit is given a row. Each variable is given a column. [You could of course do it the other way, but as in many things you'll find that you save time and headaches by being consistent with convention.] The scores (and/or in some cases, the values) the variables take on for each case are entered in the cells. [Draw on board.]

1.4 Summarizing data: displays

- So we've got a group of units, and we know the values taken on by a set of variables in this set of units. Our next move is typically to say something meaningful about the group with the data at hand. [Gesture to table.] Why not just present this?

- You're not going to get any more detailed than this. But descriptions of data usually involve tradeoffs between detail and parsimony.

- So let's move to a higher level of abstraction. A frequency table.

  - Displays the frequency with which a variable takes on values in a group of units.
  - Columns with value, number of units, proportion of units
  - One row for each value.
  - And for ordinal, interval, and ratio-level data: cumulative percentile.

- When displaying frequency distributions, sometimes it makes more sense to do so graphically. Typically graphs move a bit toward parisomony at the expense of some detail. [scale on board] What information is missing from the graph that isn't in the table?

  - Another choice: **recode** the data into categories and then produce either a frequency table or a bar chart of the categories. [e.g. on board]

- When we move to interval- or ratio-level data with lots of categories, a histogram is almost always preferred to a frequency table.

1.5 Summarizing data: measures of central tendency and dispersion

- An even higher level of abtraction: measures of central tendency and measures of dispersion.
  - Measures of **central tendency** tell us about the *typical value* of the variable in a group of units.
    - mode: the value of the variable most frequently observed in the group of units (all LOMs; wait to say this): ;
    - median: the value of the smallest observation for which the cumulative percentage is 50 or greater (ordinal and up);
    - mean - the average: $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ (interval and up).
    - Note that all of these measures summarize all the observations of a variable with just one number.
  - Measures of **dispersion** typically accompany measures of central tendency. They provide a sense of the "spread" of a variable's distribution – a.k.a. the amount of *variation* in its distribution.
    - *range* (the difference between largest and smallest values);
    - *IQR* (the difference between the values at the 75%ile and 25%ile,
    - *variance* $s^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$. Note different from book, because more intuitive. The standard deviation (s.d.) is simply $\sqrt{s^2}$. The s.d. is a more informative measure of dispersion: it is the average distance of an observation from the mean. It is measured in the same units as the variable itself. What sign must the variance and thus s.d. always be? (LOM: interval or higher):
  - Furthermore, social scientists often speak *qualitatively* about the frequency distribution of a variable:
    - It may be symmetric [draw] or skewed [draw], also number of children on handout.
      - Typically the median is a better measure of central tendency for variables with skewed distributions than the mean, because it is resistant to outliers. Use age at married example.
    - It may be unimodal or bimodal [number of children]

* There are *quantitative* ways to speak about the distribution of a variable. For example the **skewness** of a distribution is typically calculated as $g_1 = \frac{1}{N \cdot s^3} \sum_{i=1}^{N} (y_i - \overline{y})^3$ This statistic will be zero in a perfectly symmetric distribution. It will be *negative* if observations below the mean tend to be farther way from it than observations above the mean/data are skewed left/long left "tail"). If will be *positive* if data are skewed right (vice versa/long right "tail"). You rarely see this statistic used in political science, but it will pop out [handout].

2 Lecture 2

[retrieve questionnaires.]

2.1 The Normal distribution

- – One more thing: the Normal distribution.
    - ∗ Bell shaped.
    - ∗ Unimodal, symmetric.
    - ∗ Many variables, empirically, are distributed in a way that approximates the Normal distribution. E.g.: height.
    - ∗ When this is the case, we can use a rule of thumb called Tchebysheff's Theorem to describe the distribution:
        - · the interval $\overline{y} \pm s$ contains about 68% of the observations
        - · the interval $\overline{y} \pm 2s$ contains about 95% of the observations; and
        - · the interval $\overline{y} \pm 3s$ contains about all of the observations.
    - ∗ Example: the distribution of the height of American females approximates the Normal distribution. Its mean, $\overline{y}$, is 63.5 and its s.d., s = 3. Your cousin's best friend's new roommate is a woman who is 69.5 inches tall. She is taller than what proportion of females in the American population? {draw on board}
    - ∗ Because the distribution of the mean of independently drawn observations from any population tends toward the Normal when the sample size gets large, we think a lot about the Normal. This fact also creates lots of empirical Normal distributions. More to come.
    - ∗ However, do NOT make the mistake of assuming all distributions are Normal. Not e.g.: internet usage. Not e.g.: number of children.
  - – Inevitably, we find ourselves talking about how means and dispersion change over time and/or comparing groups. This begins to invite multivariate statistics.
  - – But for now, that's it for descriptive statistics with regard to one variable. Pretty simple

stuff, really. But social scientists spend a lot of time discussing and displaying descriptive statistics. Pay attention to talks in the department; you'll see.

  - Note what we haven't done: talked about samples versus populations. We've taken the data as given, and have resisted making inferences to a population.
  - That's our next big step. But to get there, we need to ascend through the basics of probability theory.

2.2 Probability theory: the logic

- We are all familiar with the process of moving from populations to samples. I have a 52-card deck that contains 13 spades. What's the probability of drawing a (sample) spade at random from a perfectly shuffled deck?
  - Obviously, it's $1/4$.
  - We know this because we know the distribution of spades in the "population," and we have been precise about the sampling process.
- Well, we're now going to make the trip in reverse.
  - Now, we know the sample. We know its central tendency and its dispersion.
  - We make precise (and justified) assumptions about the sampling process.
  - These allow us to make very good guesses about the population's central tendency and dispersion using the sample's central tendency and dispersion.
  - We need the tools of the population-to-sample process to understand the sample-to-population process.
  - So let's learn them!

2.3 A probabilistic model for an experiment

- In probability theory, we use the term **experiment** to refer to *the process by which an observation is made*. An **observation** is a quantity of interest:
  - the price of a stock

- the number of experimental subjects who choose "A" instead of "B"
- the proportion of Pew survey respondents who approve of Obama's job as president
- the proportion of Gallup ""
- in this usage, experiments don't just happen in labs, but the term is helpful because it gets us thinking about social processes in the experimental context.

- Experiments have one or more outcomes called **events**.
  - Experiment: gubernatorial election in an imaginary country that has 101 voters; majority rule, everyone votes for Candidate A or B.
  - Possible events (list): a. Candidate A wins. b. Candidate B wins. c. Candidate A wins with 76 votes. d. Candidate A wins with 56 votes. e. Candidate A wins in a landslide ($\equiv$ 67 voters or greater). How about some others? (In this example, we do not care who voted for whom; we care only about the aggregate result.)
  - Note that the first event can be decomposed into [how many] other events?
  - How about the last event?
  - a, b and e are **compound** events. By contrast, c. and d. are **simple** events. Let's formalize this. Because certain concepts from set theory will be helpful for expressing relationships amoung events, we will also associate a distinct point–a **sample point**–with each simple event. We use E_i to refer to the simple event or sample point i. Draw on board: circle S with dots denoting Es.
  - The **sample space** S associated with an experiment is the set consisting of all possible sample points. How many simple events are in S for the hypothetical election ? (102; list on board with a table)
  - In set notation, we write $S = \{E_1, E_2, ...E_{102}\}$, where E_i denotes candidate A's number of votes plus one.
  - This sample space consists of a countable (finite) number of sample points. By definition, it is a **discrete sample space.**
  - Simple events / sample points are mutually exclusive.

- By contrast, compound events are sets of sample points. The compound event A, "A wins by landslide" occurs if and only if one of the events $E_{68}...E_{102}$ occurs. Thus $A = \{E_{68}...E_{102}\}$.
  - A simple event E_i is included in compound event A if and only if A occurs whenever E_i occurs.
  - Now we can be more specific about what an **event** is: it is a collection of sample points (a subset of S).
  - We can also consider S an event in itself. It occurs whenever the experiment occurs.

- We're now ready to construct a probabilistic model for an experiment with a discrete sample space. We do so by *assigning a number, $P(A)$, to each event A* in S. Of course, we can't do this willy-nilly. We do so that *three axioms of probability* hold:
  - Axiom 1: $P(A) \geq 0$.
  - Axiom 2: $P(S) = 1$. S occurs every time the experiment is performed.
  - Axiom 3: For any sequence of pairwise mutually exclusive events $A_1, A_2, A_3,...A_n$, it must be the case that $\text{P}(A_1 \cup A_2 \cup A_3... \cup A_n) = \sum_{i=1}^{n} P(A_i)$. That is, the relevative frequency of the union of these mutually exclusive events (that is, an event made up of these events) equals the sum of their relative frequences.
  - When we assign numbers $P(A)$ to events in this way, the numbers are defined as the **probabilities of A.**
  - There are deep-thought ideas about what probability actually means. But the most intuitive way to think about probability of an event A is the proportion of the time event A would occur if we were to repeat the experiment, in the exact same way, infinitely many times.

2.4 Calculating the Probability of an Event of Interest: overview

- The three axioms of probability put bounds on the probabilities we assign to events of interest. Now we need to talk about how to assign probabilities in a rigorous way. Your text

discusses two such methods; often either can be used to assign probabilities in a given situation. The first is the *sample point* method. We won't be discussing this here in class; but it incorporates many of the techniques you learned regarding probability in Math Camp. The second is the *event composition* method. Because this method illustrates many of the laws of probability that you will use throughout your quantitative analysis classes, we will spend a bit of time reviewing this method.

2.5 The event-composition method

- In short, the *event-composition* method proceeds by decomposing and composing event A into *unions* and *intersections* of events with conveniently calculated probabilities. Four tools help you do this:
  - * the definitions of conditional probability and independence
    * multiplicative and additive laws
    * the probability of an event and its complement
    * the law of total probability and Bayes' Rule
- Before proceeding, recall that
  - $A \cap B$ "A intersection B" is the compound event where *both* A and B happen. (AND)
  - $A \cup B$ "A union B" is the compound event where either A or B happens, or both. (OR)
  - [Diagram on board]
- We require definitions of [put these on left-hand board for easy referral]:
  - **conditional probability**: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
    * Note that this is a *definition* that corresponds with notions of probability. As such it cannot be proven, but it can been shown that the definition corresponds with commonsense notions of probability. It has the intuitive meaning that $P(A|B)$ is the probability that both A and B occur given as a proportion of the probability that B occurs. The same is true for the following definition...

– **independence**: For two events A and B, if *any one* of the following conditions holds:

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$
$$P(A \cap B) = P(A)P(B)$$

, then A and B are said to be *independent* events. If *none* of these three conditions holds, then the events are said to be *dependent*.

- We also require two *laws.* The first is about the *intersection* of two or more events; the second is about the *union* of two or more events.

  – **multiplicative law** (*intersection*):

    * $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$ OR
    * $P(A \cap B) = P(A)P(B)$ if $A,\ B$ are independent events.
    * Note this law follows directly from our definition of conditional probability.
    * can be extended to 3 events:
    *

$$\begin{aligned} P(A \cap B \cap C) &= P((A \cap B) \cap C) \text{ [associative property]} \\ &= P(A \cap B)P(C|A \cap B). \text{[applying the multiplicative law once]} \\ &= P(A)P(B|A)P(C|A \cap B). \text{[applying the law again]} \end{aligned} \quad (1)$$

    * can be extended to k events: $P(A_1 \cap A_2 \cap ... \cap A_k) = P(A_1)P(A_2|A_1)...P(A_k|A_1 \cap A_2 \cap ... \cap A_{k-1})$.

  – **additive law** (*union*):

    * $P(A \cup B) = P(A) + P(B) - P(A \cap B)$,
    * where if A, B mutually exclusive, $P(A \cap B) = 0$.
    * why? Look at Venn Diagram. We're avoiding "double-counting" the intersection of events A and B.:ftbpF3.8873in2.2286in0ptFigure

- A proof:

$$\begin{aligned} A \cup B &= A \cup (\tilde{} A \cap B)\text{, and } A, (\tilde{} A \cap B) \text{ are mutually exclusive events. Also,} \\ B &= (\tilde{} A \cap B) \cup (A \cap B)\text{, and } (\tilde{} A \cap B), (A \cap B) \text{ are mutually exclusive events. Then} \\ P(A \cup B) &= P(A) + P(\tilde{} A \cap B) \text{ and } P(B) = P(\tilde{} A \cap B) + P(A \cap B) \text{ [by Axiom 3] and thus} \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B).\blacksquare \end{aligned}$$

- it's helpful to remember that the probability of **complementary events** is such that $P(A) = 1 - P(\sim A)$. Proof is easy:

$$\begin{aligned} S &= A \cup \tilde{} A \text{ (since } A \text{ is an event in sample space } S) \\ P(S) &= P(A) + P(\tilde{} A) \text{ (by Axiom 3, since } A, \tilde{} A \text{ mutually exclusive)} \\ 1 &= P(A) + P(\tilde{} A) \text{ (by Axiom 2)} \\ P(A) &= 1 - P(\sim A).\blacksquare \end{aligned}$$

- So let's run through an example using the event composition method.
  - I randomly assign a group of 16 students into 3 teams of 6, 5 and 5 students. 11 of the students are male.
  - What is $P(A)$ =the probability that the team of six students (call this "Team 1") is entirely male?
  - First, note that the event "all members of Team 1 are male" is equivalent to the event "the first six assigned students are all male."
    - Let's decompose this event into simpler events A through F:
    - A: the first student picked is male.
    - B: the second student picked is male.
    - F: the sixth student picked is male.
  - The event of interest occurs if and only if all events A through F occur. It is thus the intersection of all these events: $A \cap B \cap C \cap D \cap E \cap F$. We therefore want to find $P(A \cap B \cap C \cap D \cap E \cap F)$.

- What should we do?
  - Re-write using multiplicative law:

$$\begin{aligned} P(A_1 \cap A_2 \cap ... \cap A_k) &= P(A_1)P(A_2|A_1)...P(A_k|A_1 \cap A_2 \cap ... \cap A_{k-1}), \text{ so} \\ P(A_1 \cap A_2 \cap ... \cap A_6) &= P(A_1)P(A_2|A_1)...P(A_k|A_1 \cap A_2 \cap ... \cap A_5) \end{aligned}$$

- What is the probability that the first student picked is male? $\frac{11}{16}$.

$$P(A_1 \cap A_2 \cap ... \cap A_6) = \frac{11}{16}P(A_2|A_1)...P(A_k|A_1 \cap A_2 \cap ... \cap A_5)$$

- Given A1, what's the probability that the second student picked is male? $\frac{10}{15}$.

$$P(A_1 \cap A_2 \cap ... \cap A_6) = \frac{11}{16} \cdot \frac{10}{15} \cdot ...P(A_6|A_1 \cap A_2 \cap ... \cap A_5)$$

- And so forth:

$$P(A_1 \cap A_2 \cap ... \cap A_6) = \frac{11}{16} \cdot \frac{10}{15} \cdot \frac{9}{14} \cdot \frac{8}{13} \cdot \frac{7}{12} \cdot \frac{6}{11} = \frac{3}{52}$$

- In many instances (including exams), we are happy to write this in factor notation. How might we do this?
  - Note that the product of the numerators is $11 \cdot 10 \cdot ... \cdot 2 \cdot 1$ with the final five multiplicands "shaved off."
  - This is just $\frac{11 \cdot 10 \cdot ... \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$, or $\frac{11!}{5!}$.
  - Similarly, the denominator can be written $\frac{10!}{16!}$.
  - So another way to write this is:

$$P(A_1 \cap A_2 \cap ... \cap A_6) = \frac{11!}{5!} \cdot \frac{10!}{16!}$$

2.6 The law of total probability and Bayes' Rule

- finally, the *law of total probability* and *Bayes' Rule* can be helpful in the event-composition approach to assigning probabilities to events.

- recall that we are working with a discrete sample space S, composed of simple events.
  - we can of course also view S as a union of k mutually exclusive subsets:
    - $S = B_1 \cup B_2 \cup ... \cup B_k$;
    - $B_i \cap B_j = \varnothing, \forall i \neq j$.
    - A collection of sets $\{B_1, B_2 ... B_k\}$ such that (1) their union is equivalent to S and (2) are themselves mutually exclusive is said to be a **partition** of S.
    - If A is a subset of S, it may be **decomposed** as the union of its intersections with each of the partitions of S as follows: $A = (A \cap B_1) \cup (A \cap B_2) \cup ... (A \cap B_k)$. Draw Figure 2.12 {p. 71} on board as illustration.
    - ftbpF4.782in3.3931in0ptintersection.tif
  - The partitioning and decomposing notions yield the following:
    - If the collection $\{B_1, B_2 ... B_k\}$ is a partition of S such that $P(B_i) > 0$ for $i = 1, 2, ...k$, then the **Law of Total Probability** states:

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

    - Proof {brief}:
      - If $i \neq j$, the intersections $A \cap B_i$ and $A \cap B_j$ do not overlap. Put formally, their intersection is the empty set, because

$$\begin{aligned} (A \cap B_i) \cap (A \cap B_j) &= A \cap (B_i \cap B_j) \quad \text{[distributive law]} \\ &= A \cap \oslash \\ &= \oslash. \end{aligned}$$

        The events $A \cap B_i$ and $A \cap B_j$ are thus mutually exclusive.

- Because A is the union of all these mutually exclusive events, we can write

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + ... + P(A \cap B_k) \quad \text{[additive law]} \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + ...P(A|B_k)P(B_k) \text{ [muliplicative law]} \\ &= \sum_{i=1}^{k} P(A|B_i)P(B_i) \quad \blacksquare. \end{aligned}$$

- Why do we care? Because there are lots of instances where it's easier to calculate $P(A|B_i)$ than $P(A)$. In other cases, we intrinsically care about $P(A|B_i)$. I'll show you this in a minute, but first let's derive one additional important result.

* If the collection $\{B_1, B_2...B_k\}$ is a partition of S such that $P(B_i) > 0$ for $i = 1, 2, ...k$, then

$$\begin{aligned} P(B_j|A) &= \frac{P(A \cap B_j)}{P(A)} \text{ [definition of conditional probability]} \\ &= \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)} \text{ [definition again (num), law of total probability (denom)]} \end{aligned}$$

* This is **Bayes' Rule**.

* Let's write this in a simple case, where there are only two partitions in the sample space:

itbphF4.0938in2.4641in0inFigure

* We then have:

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)}$$

* or more simply (writing B_1 as simply B):

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\tilde{}B)P(\tilde{}B)}$$

* So classic example (2.124, page 73):

$$\begin{aligned} P(Dem|Favor) &= \frac{P(Favor|Dem)P(Dem)}{P(Favor|Dem)P(Dem) + P(Favor|REP)P(REP)} \\ &= \frac{(.7)\,(.6)}{(.7)\,(.6) + (.3)\,(.4)} \\ &= 7/9 \\ &= .\bar{7} \end{aligned}$$

* Note our process of determining the probability $P(Dem|Favor)$. We used the definition of conditional probability (which undergirds Bayes' Rule) to decompose the set (DEM $\cap$ FAVOR).

– Another classic example (check if there's time) Ex. 2.125, p. 73.

* Begin by having students write $P\,(Disease|Pos)$ according to the definition of Bayes' Rule.
* What information do we need to know?
  · $P\,(Pos|Disease) = .90$
  · $P\,(\tilde{}Pos|\tilde{}Disease) = .90$
  · $P\,(Disease) = .01$
* And so:

$$\begin{aligned} P\,(Disease|Pos) &= \frac{P(Pos|Disease)P(Disease)}{P(Pos|Disease)P(Disease) + P(Pos|\tilde{}Disease)P(\tilde{}Disease)} \\ &= \frac{(.9)(.01)}{(.9)(.01) + (.10)(.99)} \\ &= \frac{.009}{.009 + .099} = \frac{1}{12} \end{aligned}$$

* Put this in perspective:
  · What is the probability of a false positive? $P\,(Pos|\tilde{}Disease) = .10 = \frac{1}{10}$
  · What is the probability of a false negative? $P\,(\tilde{}Pos|Disease) = .10 = \frac{1}{10}$
* What affects the reliability of this test?

3 Lecture 3

- We've now provided an intellectual foundation for our understanding of probability. We've conceived of probabalistic events in the context of an **experiment** whose results are–in their most basic form–**simple events**. The set of all possible events for a given experiment is its **sample space.**
  - We then specified 3 axioms governing the assignment of probabilities to these events;
  - And then discussed four tools by decompose and compose events of interest into events with conveniently calculated probabilities.
- Before moving on, I want to formalize one additional tool that we use without thinking about it–which is *assigning probabilities in a sample space consisting of equiprobable events*:
  - Last time we had the example of assigning 16 students into 3 teams of 6, 5 and 5 students, with 11 of the students male.
    - As you may recall, we began by considering the probability that the first student picked is male. What is it? $\frac{11}{16}$. Why?
  - Here's how to formalize our intuition.
  - Consider a sample space S consisting of *n equiprobable* simple events E such that $P(E_1) = P(E_2) = ...P(E_n)$.
    - Recall that we define an A as a subset of S – that is, some subset of sample points that make up S.
    - Then $P(A) = \frac{|A|}{n}$,.where the notation $|A|$ stands for the number of elements in the set A – which in this context is the number of sample points in A.
  - Can you see how we get this from the axioms of probability?
    - First let's show that for all events E to be equiprobable, it must be that

$$P(E_i) = \frac{1}{n} \text{ for } i = \{1, 2...n\}$$

    - This is because

* $P(S) = P(E_1 \cup E_2 \cup ... \cup E_n) = P(E_1) + P(E_2) + ... + P(E_n) = 1$ [by Axiom 3 and then Axiom 2]
* $P(E_1) + P(E_1) + ... + P(E_1) = nP(E_1) = 1$ [definition of equiprobability]
* $P(E_1) = \frac{1}{n}$; and similarly $P(E_i) = \frac{1}{n}$ for $i = \{1, 2...n\}$
* Finally, since

$$\begin{aligned} P(A) &= \sum_{E_i \in A} P(E_i) \text{ [Axiom 3 again]} \\ P(A) &= \sum_{E_i \in A} \frac{1}{n} = \frac{|A|}{n} \blacksquare. \end{aligned}$$

* [For reference:]
  - Axiom 2: $P(S) = 1$. S occurs every time the experiment is performed.
  - Axiom 3: For any sequence of pairwise mutually exclusive events A_1, A_2, A_3,...A_n, it must be the case that P($A_1 \cup A_2 \cup A_3 ... \cup A_n$) $= \sum_{i=1}^{n} P(A_i)$.

– The point here is that even our intuitive basic notion of assigning probabilities to events is derived from the axioms of probability in logical ways.

3.1 The notion of a random variable

- Now let's discuss a particular kind of experiment: one in which the events of interest are numerical–that is, they are identfied in a meaningful way by numbers.
  – Numerical events of interest to political scientists might be the number of seats the Republican Party will hold in the House of Representatives after a midterm election, or the number of signatories to a treaty.
  – In the language of an experiment, we assign a *real number* to each point in the sample space.
  – Call this number the variable Y.
    * Think back: what's a *variable*? [Make them look this up]: Recall that a variable is a logical grouping of attributes. Variables take on values that are exhaustive and mutually exclusive.

- Each sample point can only take on one value of Y. But the same value may be assigned to multiple sample points.
  - [Draw a circle with S, the sample space. Draw another circle, Y, the variable. Draw lines from several points in S to one point in Y.]
- What do we call such a mapping? A *function*. Thus the variable Y is a *function* of the sample points in S.
- Recall that a function is a mathematical relation assigning each element of one set (the source) to one and only one element of another set (the target).
- In this case, the function's source is S and its target is Y. We can write $f : S \to Y$. This function (and by extension, Y) is defined as a **random variable**. Keep this in mind: whenever we talk about a random variable, *we are really talking about a function* that maps each simple event in a sample space S to a (meaningful) number.
- We write a random variable with a capital letter: Y.
  - Consequently, we write "the probability that $Y = 0$" as $P(Y = 0)$, and so forth.
  - Typically, we refer to any observed or hypothetical value of Y with a lowercase letter denoting the value. So we write $P(Y = y)$, for example.
  - We will still talk about the event of interest A, but we will give it a number, a. The event A thus $\equiv$\{all sample points such that $Y = a$\}.
- More about random variables very shortly. But first:

3.2 Quick detour: The notion of a random *sample*

- In our subsequent discussions of probability, we will often invoke the notion of a **random sample**. Let's take a moment to place this idea in the language of our probabilistic model of an experiment.
- In this framework:
  - our experiment is the drawing of a **sample** (the units selected for analysis) from a **population** (the group of units about which we wish to make inferences)

- the **design** of our experiment is the method of sampling.
    - one big decision we make, for example, is whether to sample *with replacement* (units selected for the sample are "placed back into the population" and may therefore again be sampled) or *without replacement* (units selected are set aside and cannot be selected again). Keep these terms in mind; we'll revisit them later.
- the most common way to draw a sample is called *random sampling.* Let N and n represent the numbers of elements in the population and sample, respectively. In this context, a total of [how many?] $\binom{N}{n}$ [the binomial coefficient] different samples without replacement (how many is that? $\frac{N!}{n!(N-n)!}$) may be drawn. If sampling is conducted in such a way that each of these samples has an equal probability of being selected, then we have engaged in *random sampling*, and the result is said to be a *random sample*.

3.3 Back to random variables: probability distributions

- We'll begin our discussion of random variables by focusing on those that are **discrete**. A random variable Y is discrete if it can take on only a finite or countably infinite (a one-to-one correspondence with the integers) number of distinct values.
- In making inferences from samples to populations,
    - we need to know the probability of having observed a particular event.
    - events of interest are frequently numerical events that correspond to values y of discrete random variables Y.
        - An example of such an event might be that 45 out of 100 people we surveyed said they intended to vote for Mitt Romney.
    - In other words, we need to know $P(Y = y)$ for all the values the RV Y can take on.
    - This collection of probabilities is called the **probability distribution** of the RV Y.
    - As an example, consider the experiment: roll a pair of six-sided dice and record the sum of the pips on their faces.
    - The sample space S consists of 36 simple events.

* Draw a square, label it S, divide into grid of 6 x 6.

– Define the random variable Y = the sum of the pips appearing on the faces of a random roll of a pair of six-sided dice.

* Draw another square, label it Y, and map a few of the sample points in S to Y.
* Formally, $P(Y = y) = \sum_{E_i:Y(E_i)=y} P(E_i)$. The probability $Y = y$ is defined as the sum of the probabilities of the sample points in S assigned the value y. Sometimes we write this $p(y)$.
* It its most primitive form, Y's probability distribution is a *table* in which each y is listed alongside $P(Y = y)$ But it may also be a *formula* mapping each y to its $P(Y = y)$, or a *graph* doing the same thing.
* Table: (entitled "The Probability Distribution of the RV Y")

y	# of sample points	$P(Y = y)$
2	1	$\frac{1}{36}$
3	2	$\frac{2}{36}$
4	3	$\frac{3}{36}$
5	4	$\frac{4}{36}$
6	5	$\frac{5}{36}$
7	6	$\frac{6}{36}$
8	5	$\frac{5}{36}$
9	4	$\frac{4}{36}$
10	3	$\frac{3}{36}$
11	2	$\frac{2}{36}$
12	1	$\frac{1}{36}$
totals	**36**	**1**

- – * Graph: [draw on board]
  * A function (here's one way, there may be others/better). Here, $p(y)$ is known as a **probability function**.

$$P(Y = y) = p(y) = \frac{6 - |7 - y|}{36}, y = \{1, 2, ...6\}.$$

- The probability function of a discrete random variable is also called its **probability mass function, or PMF.** The "mass of a random variable at y" is the PMF evaluated at y, or $p(y)$.
- The probabilities associated with distinct values of y sum to 1, and in fact they always do, as the second axiom of probability requires.

3.4 Random variables: expected values

- The probability distribution of an RV is a *theoretical model* for the empirical distribution of data associated with a real population. If we were to roll a pair of dice over and over again, recording the sum of the faces each time, the empirical distribution would look very much like the theoretical probability distribution of Y we just specified.
- As we do with empirical data, we can describe a random variable by talking about its central tendency and dispersion. (At what level of measurement are we implicitly working here? Interval or higher.) So we will be concerned about the mean and variance of a random variable.
- We specify and manipulate formulas describing random variables using the *expectations operator*, which is defined as follows:
  - Let Y be a discrete RV with the probability function $p(y)$. The **expected value** of Y is written $E(Y)$, and defined to be:

$$E(Y) \equiv \sum_{y} yp(y).$$

  - Note that this is each possible value of Y times the probability that Y takes on the value y, i.e., $p(y)$, summed up over all y. [Calculate this value for the two dice example.]
  - So the expectations operator tells us to consider all the possible values of a RV, multiply each of these values by their probability of occurrence, and to sum up these products.
  - The expected value is the way we talk about the central tendency of a random variable with a theoretical probability distribution. It is equivalent to the idea of the mean of an empirical frequency distribution.

- Now we take what is a very powerful step as we move into inferential statistics. Recall that the probability distribution of an RV is a *theoretical model* for the empirical distribution of data associated with a real population. If this theoretical model is accurate, then $E(Y) = \mu$, the *population mean*.

$$E(Y) \equiv \sum_y yp(y) = \mu.$$

- Here, μ is a *parameter*: a characteristic of the distribution of Y in the population that we never actually observe but about which we are often keenly interested. It is the first of many such parameters we will encounter in this class.

- We are often interested in the expected value of *functions of random variables.* (You'll see why in a minute.) Consider as usual the discrete RV, Y with probability function $p(y)$. Now consider any real-valued function of Y, $g(Y)$. Then the expected value of $g(Y)$is given by:

$$E[g(Y)] = \sum_y g(y)p(y).$$

- That is, the expected value of a function of a RV is given by:
  - evaluating the function for each value of Y
  - multiplying it by the probability that $Y = y$
  - and summing up over all possible values of Y.

- Now this is not a definition, but rather can be proven based upon the definition of $E(Y)$ .Proof:
  - Consider a discrete RV Y taking on a finite number, n, of values $y_1, y_2, ...y_n$.
  - Suppose $g(y)$ takes on m different values $g_1, g_2...g_m, m \leq n$.
    - We're being this picky because it's of course possible that $g()$ is not one-to-one; that is, it may take on the same value for multiple y's.
  - Note that $g(Y)$ is itself a random variable (by definition of RV). It is a function mapping the sample space Y to the reals.

– So we can define a new probability function, p-star, for g.

– Write the probability that g takes on a value g_i as

$$\begin{aligned} p^*(g_i) &= P[g(Y) = g_i] \\ &= \sum_{y_j : g(y_j) = g_i} p(y_j), \text{ for } i = 1, 2, ...m \end{aligned}$$

– where $y_j : g(y_j) = g_i$ means "all y_j such that g equals g_i when evaluated at y_j."

– Now the definition of expected value tells us that:

$$E[g(Y)] = \sum_{i=1}^{m} g_i \, p^*(g_i)$$

– (Note that we're doing the same thing as before:

  * Taking each possible value of g, that is all the g_i's
  * Multiplying it by its associated probability (which we've defined in this case as $p^*(g_i)$)
  * And summing up these products.)

– Moving on with proof:

$$\begin{aligned} &= \sum_{i=1}^{m} g_i \left\{ \sum_{y_j : g(y_j) = g_i} p(y_j) \right\} \text{ [substituting]} \\ &= \sum_{i=1}^{m} \left\{ \sum_{y_j : g(y_j) = g_i} g_i \, p(y_j) \right\} \text{ [can move } g_i \text{ inside because of nested summations]} \\ &= \sum_{j=1}^{n} g(y_j) p(y_j) . \text{[since } g_i = g(y_j) \text{ for any } y_j; \text{ we change index to signify we're now interested in all } n \text{ values that } Y \text{ can take on]} \\ &= \sum_{y} g(y) p(y). \text{ [writing more simply]} \end{aligned}$$

– Again, the intuition behind $E[g(Y)] = \sum_y g(y)p(y)$ is straightforward. We proceed by:

  * evaluating the function for each value of Y
  * multiplying it by the probability that $Y = y$

* and summing up over all possible values of Y.

- We can now define the variance of a random variable Y. Recall before that we defined the variance of an empirical variable as $s^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$.
  - Well, the variance of a random variable is how it varies about its mean:

$$VAR(Y) \equiv E[(Y - E(Y))^2]$$

  - And if the RV Y accurately describes the population distribution then

$$VAR(Y) = E[(Y - \mu)^2].$$

  - The standard deviation of Y is $+\sqrt{VAR(Y)}$.
  - And (again assuming that our RV variable is an accurate theoretical model of the world) then $VAR(Y) = \sigma^2$, the population variance, with $\sqrt[+]{\sigma^2} = \sigma$ the population standard deviation.
  - Do Example 3.2 (p. 94) on the board.

3.5 Some helpful results

- Go over "Some helpful results about the math of expectations for discrete RVs" handout. Proofs:
- To show $E(c) = c$:
  - Consider the function $g(Y) \equiv c$. By the expectation of a function of a random variable theorem, $E(c) = \sum_y cp(y) = c\sum_y p(y)$. But by Axiom 2, $\sum_y p(y) = 1$. Hence $E(c) = c(1) = c$. ■.
- To show $E[cg(Y)] = cE[g(Y)]$:

– By the expectation of a function of a random variable theorem,

$$\begin{aligned} E[cg(Y)] &= \sum_y cg(y)p(y) \\ &= c\sum_y g(y)p(y) \\ &= cE[g(Y)] \quad \blacksquare. \end{aligned}$$

- To show that we can distribute expectations, consider the case where $k = 2$:

$$\begin{aligned} E[g_1(Y) + g_2(Y)] &= \sum_y [g_1(y) + g_2(y)]p(y) \quad \text{[since } g_1(Y) + g_2(Y) \text{ is a function of } Y\text{]} \\ &= \sum_y [g_1(y)p(y)] + \sum_y [g_2(y)p(y)] \quad \text{[distributing summations]} \\ &= E[g_1(Y)] + E[g_2(Y)] \quad \text{[by definition of the expectations operator]} \\ & \quad \blacksquare. \end{aligned}$$

- All of these results help us re-write the formula for the population variance in a very helpful way (proof is on handout).

- [Emphasize: we always treat parameters as constants when applying the expectations operator. They are invariant.]

3.6 Three theoretical probability models

- As illustrations, we will discuss three of the standard discrete probability distributions:

  – the **Bernoulli**

  – the **binomial**

  – the **Poisson**

3.7 The Bernoulli

- A Bernoulli experiment is the observation of an experiment consisting of one trial with two outcomes: zero or one.

- Thus the Bernoulli random variable Y takes on the values $\{0, 1\}$.

- E.g.'s:
  - a coin toss
  - whether an individual approves or disapproves of Barack Obama's performance as president
  - whether a country signs a treaty
- A Bernoulli random variable is characterized by one parameter: π, the probability of success.
- Thus its probability distribution is

$$\begin{aligned} p(y=1) &= \pi \\ p(y=0) &= 1-\pi \end{aligned}$$

- Sometimes it's convenient to write the probability (mass) function

$$p(y) = \pi^{y}(1-\pi)^{(1-y)}$$

  - Think of the exponents as "switches" which turn on and off (i.e. equal to one) depending on the value of y at which the function is being evaluated.
- For a little practice, let's show that

$$\begin{aligned} E(Y) &= \pi. \\ E(Y) &= \sum_{y} p(y)\, y \\ &= p(y=0)(0) + p(y=1)(1) \\ &= \pi. \end{aligned}$$

- How about

$$\begin{aligned} VAR(Y) &= \pi(1-\pi) \\ VAR(Y) &\equiv E[(Y-)^2] \\ &= E(Y^2) - [E(Y)]^2 \text{ [from handout; see how it's already helpful?]} \end{aligned}$$

- Note that

$$\begin{aligned} E(Y^2) &= \sum_y y^2 p(y = Y) \\ &= 0^2(1-\pi) + 1^2\pi \\ &= \pi \end{aligned}$$

- Now substitute

$$\begin{aligned} VAR(Y) &= E(Y^2) - [E(Y)]^2 \\ &= \pi - \pi^2 = \pi(1-\pi). \end{aligned}$$

- Let's talk a bit more about **parameters**. Together with the probability function, parameters determine the shape, location and spread of the distribution. A **location parameter** specifies the location in the center of the distribution. As we change a location parameter we shift the PMF to the left or right. Its empirical referent is, of course, the mean. A **scale parameter** specifies the spread (or scale) of the distribution around its central location. Its empirical referent is the standard deviation. What's notable about the Bernoulli (and, as you will see, the Binomial and the Poisson) is that they only have location parameters; their spread and their location are tied together.

3.8 The Binomial

- A binomial experiment is the observation of a sequence of identical and independent Bernoulli trials. Specifically:

- A fixed number, n, of trials with one of two outcomes: success S or failure F.
- As with the Bernoulli, the probability of success on any single trial is π. The probability of failure is thus $1 - \pi$.
- The trials are independent.
- The random variable of interest is Y, the # of successes observed during the n trials.

- E.g.'s:
  - # of heads observed in a certain # of coin tosses
  - # of people approving of Barack Obama out of a certain # of people
  - # of countries signing a treaty out of n eligible countries

- Let's find the binomial probability distribution. Recall that a probability distribution (table, graph, formula) tells us with what probability our RV of interest, Y, takes on all possible values y, i.e. $P(Y = y)$, or simply $p(y)$.

- In the context of a probability model, we consider the event "the number of successes equals y," or $Y = y$, to be our "event of interest."

- Let's assign a probability $p(y)$ to each of these events.

- One such event might be $SSFSFFFSFS...FS$, where there are n such positions.

- If we were to count the S's and F's, we might see

$$S_1SSSS...SSS_y \qquad F_1FF....FF_{n-y}$$

- In this context, the number of S's equals what? y. And so the number of F's equals what? $n - y$.

- Think about our laws of probability. The event of interest we have witnessed is the **intersection** of n simple events: $S_1 \cap S_2 \cap ... \cap S_y \cap F_1 \cap F_2 ... \cap F_{n-y}$. These are **independent** events (by definition of the binomial experiment). So $P(S_1 \cap S_2 \cap ... \cap S_y \cap F_1 \cap F_2 ... \cap F_{n-y})$ simply equals $P(S_1)P(S_2)...P(S_y)P(F_1)P(F_2)P(F_{n-y})$.

- So what's that? It's $\pi^y (1-\pi)^{n-y}$.
- But this is *not* the probability of seeing the event $Y = y$. Why? Because the event $Y = y$ could happen in many more ways than the order in which we saw it. How many different ways are there to order y S's and $n - y$ F's? It's the number of ways of choosing y elements from a total of n elements. Or "n, choose y." From math camp, you'll remember that this is equal to (ha!) the binomial coefficient $\binom{n}{y}$,or $\frac{n!}{y!(n-y)!}$.
- Thus the probability of observing $Y = y$ is

$$p(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}.$$

This can be written as $Binomial(\pi; n)$

- In the parlance of probability distributions, the binomial has two **parameters**: π and n.
- So (very quickly). 9 students in the class, 5 male. I pick six students at random with replacement. What's the chance that I pick the same number of males and females?
  - Call 'success' a female. $\pi = \frac{4}{9}$.
  - We have # of trials $n = 6$. We have # of successes as $y = 3$.
  - We thus evaluate the probability distribution for $Bi(\frac{4}{9}; 6)$ at $y = 3$:

$$\begin{aligned} p(Y &= 3) = \frac{6!}{3!(6-3)!}\left(\frac{4}{9}\right)^3\left(1-\frac{4}{9}\right)^{6-3} \\ &\approx .30 \end{aligned}$$

- Let's say I turned this question around a bit. I draw six students at random (with replacement). On average, how many females will I pick? And how much will this number vary over repeated draws of six?
- Answering questions like this requires a formula for the expected value and the variance of the Bernoulli random variable Y.

- And in fact, we can prove that if Y is a binomial random variable,

$$E(Y) = n\pi = \mu \text{ and } VAR(Y) = n\pi(1-\pi) = \mu(1-\pi) = \sigma^2.$$

- We'll omit the proofs in class because we need to move on. But the proofs (pp. 107-108) are informative.

- Thus $E(Y|p = \frac{4}{9}; n = 6) = 6 \times \frac{4}{9} = 2.\overline{6}$. $VAR(Y) \approx 2.7\left(1 - \frac{4}{9}\right) = 1.5$. The s.d. is $\sqrt{1.5} \approx 1.2$. The typical # of females I will draw in a sample of 6 is 2.7, but the typical value will fall 1.2 females away from 2.7.

- In practice, Binomials are not conveniently calculated. We typically use statistical software to calculate the probability that the Binomial random variable Y takes on some value y. In Stata, we type `.di binomialp`(n, y, π) to get $P(Y = y)$.

4 Lecture 4

4.1 The Poisson

- A final discrete probability distribution we will examine is the Poisson.
- A Poisson experiment is the observation of a *count of events* that occur in an *interval, broadly defined* – a given space, time period, or any other dimension. It is particularly helpful when modifying relatively *infrequent* events (as more frequent events can be modeled with more generic distributions).
  - Environmental laws per Congressional session.
  - Errors per page.
  - Government shutdowns per decade.
  - Homeless shelters per census tract.
- We discuss the Poisson after the Binomial because it can actually be conceived of the Binomial experiment as the number of trials approaches infinity. (What? Let's consider:)
  - Split the interval into n subintervals, each so small that at most one event could occur in it. That is, in each subinterval a Bernoulli trial takes place:

$$\begin{aligned} P(y=1) &= \pi \\ P(y=0) &= 1-\pi \\ \text{AND } P(y>1) &= 0. \end{aligned}$$

  - In this subinterval,

$$p(y) = \pi^y (1-\pi)^{(1-y)} \text{ [Bernoulli]}$$

  - And if we have n subintervals, then

$$p(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y} \text{ [Binomial]}$$

- How many such subintervals are needed? Who knows. But we can handle this problem by making the subintervals infinitely small by taking the limit of the Binomial probability function as n goes to infinity. Our parameter of interest–the number of successes over the interval–is $n\pi$. We call this parameter $\lambda = n\pi$. So:

$$\begin{aligned} & \lim_{n\longrightarrow\infty} \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y} \\ = & \lim_{n\longrightarrow\infty} \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} \\ = & \lim_{n\longrightarrow\infty} \frac{n!}{y!(n-y)!}\frac{\lambda^y}{n^y}\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-y} \end{aligned}$$

- Noting that by definition of e,

$$\lim_{n\longrightarrow\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$$

- And so

$$\begin{aligned} = & \lim_{n\longrightarrow\infty} \frac{n!}{y!(n-y)!}\frac{\lambda^y}{n^y}e^{-\lambda}(1) \\ = & \frac{e^{-\lambda}\lambda^y}{y!}\lim_{n\longrightarrow\infty}\frac{n!}{(n-y)!n^y} \text{ [pulling out everything not related to } n\text{]} \\ = & \frac{\lambda^y}{y!}e^{-\lambda}\lim_{n\longrightarrow\infty}\frac{n(n-1)(n-2)...(n-y+1)}{n^y} \text{ [shaving off } n-y \text{ terms from numerator]} \\ = & \frac{\lambda^y}{y!}e^{-\lambda}\lim_{n\longrightarrow\infty}\frac{n}{n}\times\frac{(n-1)}{n}\times\frac{(n-2)}{n}\times...\times\frac{(n-y+1)}{n} \\ = & \frac{\lambda^y}{y!}e^{-\lambda}\lim_{n\longrightarrow\infty}1\times\left(1-\frac{1}{n}\right)\times\left(1-\frac{2}{n}\right)\times...\times\left(1-\frac{y+1}{n}\right) \\ = & \frac{\lambda^y}{y!}e^{-\lambda}(1) \end{aligned}$$

- And so

$$p(y) = \frac{\lambda^y}{y!}e^{-\lambda}.$$

- In practice, Poissons are not conveniently calculated. We typically use statistical software to calculate the probability that the Binomial random variable Y takes on some value y. In Stata, we type `.di poissonp`(λ, y) to get $P(Y = y)$.

- It is the case [proofs are in your book] that for a Poisson RV,

$$\begin{aligned} \mu &= E(Y) = \lambda \\ \sigma^2 &= VAR(Y) = \lambda. \end{aligned}$$

The take-home point is that there are many theoretical models of specific experiments. Each of them has a probability distribution $p(y)$, and each has a mean ($E(Y)$) and a variance ($VAR(Y)$). Get to know these a bit by browsing the remainder of Chapter 3. See also the inside back cover of your text for a quick summary.

4.2 Continuous random variables

- We often deal with random variables that can take on an uncountably infinite number of values. These RVs are known as *continuous* random variables.
- The reason we care: it's impossible to assign nonzero probabilities to all the uncountably infinite points on an interval while satisfying that they all sum to 1. Thus the notion of $p(y)$ from discrete world is irrelevant here. We must develop a different method to describe the probability distribution of a continuous RV.
- Let's begin by defining the cumulative distribution function (or cdf, or "distribution function") of the RV Y as $F(y)$, where

$$F(y) \equiv P(Y \leq y) \text{ for -}\infty < y < \infty.$$

ftbpF4.5637in3.2055in0ptFigure

- A cdf has the following properties:

$$\begin{aligned} F(-\infty) &\equiv \lim_{y \to \text{-}\infty} F(y) = 0 \\ F(\infty) &\equiv \lim_{y \to \infty} F(y) = 1 \\ & \quad F(y) \text{ is a nondescreasing function of } y, \text{ which means that} \\ y_1 &< y_2 \Longrightarrow F(y_1) \leq F(y_2). \end{aligned}$$

- Both discrete and continous RVs have cdfs. See your text [page 158] for an example of how to develop a cdf of an RV with a binomial distribution . A RV Y with cdf $F(y)$ is said to be **continuous** if $F(y)$ is continuous for -$\infty < y < \infty$. By contrast, the cdf's of discrete RVs are always **step** functions: they have discontinuities separating the possible values of y that they can take on.

- Remember that we talked earlier about the difficulty of assigning probabilities to points? Well with continuous RVs, we don't. In fact if Y is a continuous RV,

$$P(Y = y) = 0 \forall \text{ real numbers } y.$$

- Sounds weird, but isn't. When we move to the world of continuous RVs, the differences between values of y become infinitely small, making the chance that we see any one particular value of y zero.

  - For example, what's the probability of observing a temperature of 50.73093764 degrees Fahrenheit on October 2 in New York City? Now add 10 additional random digits to this number. Do it again. It doesn't take long to get to values for which it would be quite likely we would never see even if we were to measure the temperature on all October 2s that ever exist. And even *those* values can be made more precise.

- Instead, we get at this notion with the idea of **density**. Define the function $f(y)$ as the derivative of f:

$$f(y) \equiv \frac{dF(y)}{dy} = F'(y).$$

- Wherever the derivative exists, $f(y)$ is the *probability density function* (pdf, 'density function') for the RV Y. [NOTE: The following diagram should be labeled "The density function"

ftbpF4.181in2.6501in0ptFigure

- How to think about this intuitively:

  - Note that where the cdf is changing rapidly (has a steep slope), the density is larger. Where it is changing slowly (has a flatter slope), the density is smaller.

- Having defined $f(y) \equiv \frac{dF(y)}{dy}$, we therefore can write

$$F(y) = \int_{-\infty}^{y} f(t)dt,$$

where $f(\cdot)$ is the pdf and t is a placeholder, the variable of integration. The pdf $f(\cdot)$ has the following properties:

$$\begin{aligned} f(y) &\geq 0 \forall y, \text{-}\infty < y < \infty. \\ \int_{-\infty}^{\infty} f(y)dy &= 1. \end{aligned}$$

- Be sure to work through the examples in your book to get a sense of the math of cdf's and pdf's.

- Now what if we want to find the probability that the random variable Y falls in a certain interval, e.g. $P(a \leq Y \leq b)$? It is the area under the density function in this interval [draw]

$$\begin{aligned} P(a &< Y \leq b) = P(Y \leq b) - P(Y \leq a) \\ &= F(b) - F(a) \\ &= \int_a^b f(y)dy. \end{aligned} \tag{2}$$

- In the continuous RV case, note that $P(a < Y < b) = P(a < Y \leq b) = P(a \leq Y < b) = P(a \leq Y \leq b)$. Why?

- Note that this is not necessarily the case for discrete RVs, because for those kinds of RVs, $P(y \leq a)$ is not necessarily equal to $P(y < a)$.

4.3 Expected values for continuous random variables

- Remember that in the case of discrete RV, we wrote

$$E(Y) \equiv \sum_y yp(y) \text{ and } E[g(Y)] = \sum_y g(y)p(y)$$

The math of expectations for continuous RVs is very similar:

$$\begin{aligned} E(Y) &\equiv \int_{-\infty}^{\infty} yf(y)dy \\ E[g(Y)] &= \int_{-\infty}^{\infty} g(y)f(y)dy \end{aligned}$$

- Furthermore:

$$\begin{aligned} VAR(Y) &\equiv \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy \\ &= E(Y^2) - \mu^2. \text{ [Exercise.]} \end{aligned}$$

- Reassuringly, this is the same as in the discrete case, as we showed last time in class.

4.4 Theoretical models of continuous probability distributions

- We'll look at two continuous RVs in detail:
  - The **Uniform**
  - The **Normal**
- And three distributions related to the Normal that we will use constantly in statistical tests:
  - * the **Chi-squared** (χ^2) distribution
    * the **t-distribution**
    * the **F idstribution**

4.5 The Uniform distribution

- Consider a random variable that can take on any value in an interval between two values, and these values are equiprobable.
  - e.g., the date of the election called by a prime minister who must call an election at some point in her five-year term.

  - the length of a student essay (in words) that is required to be between 1,000 and 2,000 words.
  - the length of a Tweet, which must be between 1 and 148 characters.

- (Hold off for a moment on scrutinizing whether these values are really all equiprobable; we'll get to that in a minute.)
- (Let's also hold off for a moment at balking at the fact that these are actually examples of discrete random variables, rather than continuous RVs.)
- We can represent the density function of such an RV like this (change a and b to the thetas):ftbpF3.3831in1.970
- This flat density function represents the fact that all values in the interval are equiprobable.
- This leads to a pdf that looks like this:

$$f(y) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \theta_1 \leq y \leq \theta_2 \\ 0, & \text{elsewhere.} \end{cases}$$

- How do we get to the pdf? Simple geometry.
  - We know the rectangular area under the distribution function must equal 1.
  - We know the length of the base of the rectangle: it's $\theta_2 - \theta_1$.
  - Therefore its height h must solve $(\theta_2 - \theta_1)\, h = 1$ and thus $h = \frac{1}{\theta_2 - \theta_1}$.
- Note that when we "stretch out" the interval of a Uniform RV, its density gets smaller and smaller. Can you explain the intuition for this?
- Let's derive the cdf of the Uniform:

$$\begin{aligned} F(y) &= \int_{-\infty}^{y} f(t)dt \\ &= \int_{\theta_1}^{y} \frac{1}{\theta_2 - \theta_1} dt \\ &= \frac{t}{\theta_2 - \theta_1}\Big|_{\theta_1}^{y} = \frac{y - \theta_1}{\theta_2 - \theta_1} \end{aligned}$$

- Now let's show that $\mu = E(Y) = \frac{\theta_1 - \theta_2}{2}$.

$$\begin{aligned} E(Y) &\equiv \int y f(y)\, dy \\ &= \int_{\theta_1}^{\theta_2} y \frac{1}{\theta_2 - \theta_1} dy \\ &= \frac{1}{\theta_2 - \theta_1} \frac{y^2}{2}\Big|_{\theta_1}^{\theta_2} = \frac{(\theta_2)^2 - (\theta_1)^2}{2(\theta_2 - \theta_1)} \\ &= \frac{\theta_1 + \theta_2}{2}. \end{aligned}$$

- Of course, this is quite intutive. The expected value of a Uniform RV should be at the point that divides the interval in half!

- A final result [proof will be on your HW]:

$$\sigma^2 = VAR(Y) = \frac{(\theta_2 - \theta_1)^2}{12}$$

- OK, now let's deal with the quibbles. When we think about the examples I mentioned earlier, we know enough about them that they are not properly modeled as having all their values equiprobable. [Go through examples.]

- However, what if we began examining a social process knowing nothing about it?

  - Or more to the point, we wanted to be completely *agnostic* about how the values might be distributed?

  - Assuming that the process follows a Uniform distribution is a good way to do this. And in fact, that's exactly what we do in a lot of statistical modeling. We also do this in a lot of formal modeling when we want to be agnostic about the beliefs an actor might have about the value of something. In both cases, we call this "Uniform prior beliefs," or just a "Uniform prior."

4.6 The Normal distribution

- Many empirical distributions are closely approximated by a distribution that is:

  - symmetric
  - has positive (non-zero) probability for all possible values of y
  - and is "bell shaped": specifically it has inflection points at one standard deviation away from its mean.

- In fact, it can be shown that:
  - in repeated random samples from a population
  - the means of these samples
  - are distributed around the population mean, μ, as described by a density function that we can actually write down.
  - The proof of this phenomenon is called the Central Limit Theorem, and this density function is defined as the Normal density function.

- We won't be proving the Central Limit Theorem. In a few lectures, I will instead be giving you empirical evidence for its existence.

- So unlike the other PMFs and PDFs we've looked at so far, we won't derive this function. It requires learning a lot of mathematics that I don't think you'll find particularly relevant.

- Instead, you will take it on faith that a RV Y has a Normal probability distribution iff, for $\sigma > 0$ and $-\infty < y < \infty$, the density function of Y is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}, \text{ for } -\infty < y < \infty.$$

- Draw on board, noting μ on x-axis (axis is labeled 'y'), y-axis labeled '$f(y)$', height of curve labeled $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$.

- Note that the Normal has two parameters: μ and σ.

- As usual, if empirical values in population are distributed Normal, then

$$E(Y) = \mu \text{ and } VAR(Y) = \sigma^2.$$

- So based on what we learned earlier, what is $P(a \leq Y \leq b)$ if Y is distributed Normal?

$$P(a \leq Y \leq b) = \int_a^b f(y)dy = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy.$$

- [Shade in area of graph.]

- A closed-form expression for this interval does not exist. Evaluating requires numerical methods. Commands are available in Stata (you'll learn them).

- Typically (and you'll recall this from any statistics class!), we **standardize** a Normally distributed variable so that our units are measured in terms of standard deviations. That is we transform that normal RV Y into the standard normal RV Z :

$$Z \equiv \frac{Y - \mu}{\sigma}.$$

- The RV Z is itself distributed Normal with mean zero and standard deviation one. Generally wecan standardize any Normal variable without losing any relevant information. The original distribution is now expressed in units of the standard deviation of the original Normal RV.

- If we standardize a Normal, its pdf becomes simpler:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- We use the pdf and cdf of the standard Normal so much that we have special symbols for them:

  - We write $\phi(z)$ ("or little phi of z") to denote the pdf of the standard Normal evaluated at $Z = z$
  - and $\Phi(z) = \Pr(Z \leq z)$, (or "big phi of z") to denote the cdf of the standard Normal evaluated at $Z = z$.

4.7 Three distributions related to the standard Normal

- – There are three distributions related to the Normal that we will use constantly in statistical tests. It's not much use to go into detail about them here; I just want you to be aware that they all derived from the Normal. We will encounter them again soon.
    - * the **Chi-squared** (χ^2) distribution [where Y is the sum of the squares of a series of standard Normal RVs]
    - * the **t-distribution** [where Y is the ratio of a standard Normal RV / the square root of a chi-squared RV]
    - * the **F idstribution** [where Y is the ratio of two chi-squared RVs]