Math Camp 2012:
Probability

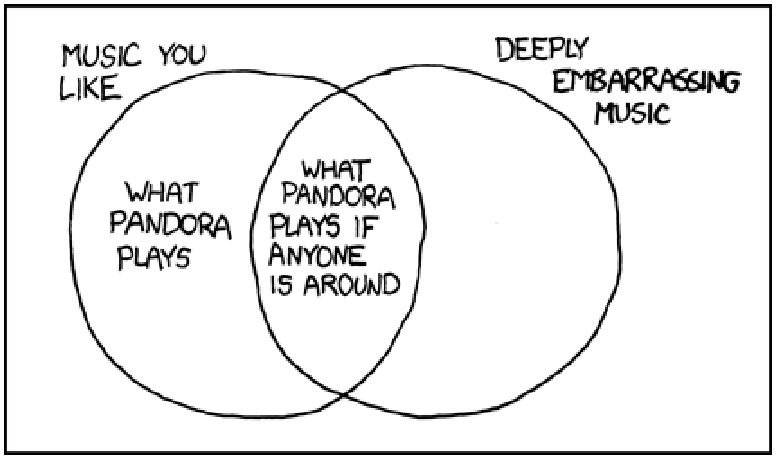
August 2012

Topics[*]:

- Counting, Sets, and Basic Probability Theory
  - Advanced Counting
  - Sets
  - Probability
  - Conditional probability and Bayes' Law
  - Independence
- Random Variables
  - Measures
  - Probability mass functions
  - Probability density functions
  - Cumulative
  - Joint distributions
  - Expected values
  - Summarizing observed data

[*] These notes were prepared by Jacob Montgomery. Much of the material and examples for this lecture are taken from Harvard "Math (P)refresher" class notes whose authors are listed here and *A Mathematical Primer for Social Statistics* by John Fox.

1 Probability


MUSIC YOU LIKE
DEEPLY EMBARRASSING MUSIC
WHAT PANDORA PLAYS
WHAT PANDORA PLAYS IF ANYONE IS AROUND

In both formal modeling and statistics classes, you will constantly interact with concepts from basic probability theory. In fact, many game theory and statistics texts will include small amounts of background information in passing. This is often helpful, but can lead some students to get confused about what distinguishes statistical inference, the basic task in both statistics and game theoretic models under uncertainty, from probability theory. So before you move forward into those areas of study, it is worth taking some time to understand probability by itself.

The goal of inference is to take some observed data or known facts and backwards induct something about the world. For instance, we might want to survey a random subset of American citizens and estimate the average attitudes of the entire American electorate. Alternatively, a game theoretic model may require actors to estimate the location of the median voter given the sequence of prior election outcomes $x = (x_1, x_2, \ldots, x_n)$ and candidate positions $y = (y_1, y_2, \ldots, y_n)$.

Probability theory is exactly the reverse. Here we *know* the basic features of the data generating process (the parameters) and want to understand what the data is likely to look like. For instance, we might have a fair coin and we want to understand the likelihood of flipping 20 heads before the first tail shows up. Obviously, most of the things you are going to be doing in graduate school will be about inference. Nonetheless, you *really* need to have a grasp of probability theory first.

Probability is essentially thinking clearly about counting. For the simplest problems, all you need to know is the number of ways that some set of outcomes X could happen versus the total number of ways things could have turned out. So to begin with, you just need to focus on getting a handle on the basic concepts of:

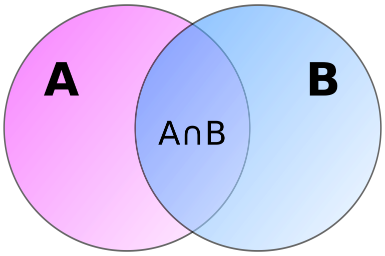
- How to count events
- How to think about and handle sets
- How counting and sets relate to the concept of "probability"
- Conditional probability, independence, and Bayes' law

1.1 Counting rules

- **Fundamental Theorem of Counting**: If there are k characteristics, each with n_k alternatives, there are $\prod_{i=1}^{k} n_k$ possible outcomes.
- We often need to count the number of ways to choose a subset from some set of possibilities. The number of outcomes depends on two characteristics of the process: does the *order* matter and is *replacement* allowed?
- If there are n objects and we select $k < n$ of them, how many different outcomes are possible?
  1. Ordered, with replacement: n^k
  2. Ordered, without replacement: $\frac{n!}{(n-k)!}$
  3. Unordered, with replacement: $\frac{(n+k-1)!}{(n-1)!k!} = \left(\begin{array}{c} n+k-1 \\ k \end{array} \right)$
  4. Unordered, without replacement: (n choose k): $\frac{n!}{(n-k)!k!} = \left(\begin{array}{c} n \\ k \end{array} \right)$
- Ordered events are sometimes referred to as permutations, while unordered events are combinations.
- In your introductory work, you will almost always be working with combinations.

1.2 Sets

- **Set**: A set is any well defined collection of elements. If x is an element of S, $x \in S$.
- Types of sets:
  1. Countably finite: a set with a finite number of elements, which can be mapped onto positive integers.
     $S = \{1, 2, 3, 4, 5, 6\}$
  2. Countably infinite: a set with an infinite number of elements, which can still be mapped onto positive integers.
     $S = \{1, \frac{1}{2}, \frac{1}{3}, \ldots\}$
  3. Uncountably infinite: a set with an infinite number of elements, which cannot be mapped onto positive integers.
     $S = \{x : x \in [0, 1]\}$
  4. Empty: a set with no elements.
     $S = \{\emptyset\}$



- Set operations:

  1. **Union**: The union of two sets A and B, $A \cup B$, is the set containing all of the elements in A or B.
  2. **Intersection**: The intersection of sets A and B, $A \cap B$, is the set containing all of the elements in both A and B.
  3. **Complement**: If set A is a subset of S, then the complement of A, denoted A^C, is the set containing all of the elements in S that are not in A.

- Properties of set operations:[†]

  1. Commutative: $A \cup B = B \cup A$, $A \cap B = B \cap A$
  2. Associative: $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$
  3. Distributive: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  4. de Morgan's laws: $(A \cup B)^C = A^C \cap B^C$, $(A \cap B)^C = A^C \cup B^C$

- **Disjointness**: Sets are disjoint when they do not intersect, such that $A \cap B = \{\emptyset\}$. A collection of sets is pairwise disjoint if, for all $i \neq j$, $A_i \cap A_j = \{\emptyset\}$. A collection of sets form a partition of set S if they are pairwise disjoint and they cover set S, such that $\bigcup_{i=1}^{k} A_i = S$.

1.3 Probability

- **Probability**: Probability is an expression of uncertainty. Modern probability theory is a way of estimating our uncertainty about some future events given specific assumed properties of the world. This is a formalization of basic human intuition about how to handle risk.

- **Sample Space**: A set or collection of all possible outcomes from some process. Outcomes in the set can be discrete elements (countable) or points along a continuous interval (uncountable).

- Examples:

  1. Discrete: the numbers on a die, the number of possible wars that could occur each year, whether a vote cast is republican or democrat.
  2. Continuous: GNP, arms spending, age.

- **Probability Distribution/Function**: A probability *function* on a sample space S is a mapping $\Pr(A)$ from events in S to the real numbers. It is just like any other function. We have some event/sample space S we have a probability space (e.g., the probability of event x happening is some number in $[0, 1]$)and we have the function that translates x into the probability space that we denote $p(x)$ or $f(x)$.

- **Axioms of Probability**: Probability functions will satisfy the following three axioms (due to Kolmogorov). Define the number $\Pr(A)$ corresponding to each event A in the sample space S such that

  1. Axiom: For any event A, $\Pr(A) \geq 0$.
  2. Axiom: $\Pr(S) = 1$

[†] These are also *very* important for understanding basic computer programming.

3. Axiom: For any sequence of disjoint events $A_1, A_2, \ldots$ (of which there may be infinitely many),
   $\Pr\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \Pr(A_i)$

- **Example**: Let's say we are flipping two coins. The sample space is $S = HH, HT, TH, TT$.

Outcome	X=x
HH	2
HT	1
TH	1
TT	0

x	p(x)
TT → 0	.25
HT, TH → 1	.50
HH → 2	.25
Sum	1.00

- **Example**: Avoiding being a sure loser – Axioms of probability and betting.

  1. A1: Rain and High above 68 degrees F tomorrow
  2. A2: Rain and High at or below 68 degrees F tomorrow
  3. A3: No Rain and High above 68 degrees F tomorrow
  4. A4: No Rain and High at or below 68 degrees F tomorrow

  – These events are disjoint and exhaustive.
  – Now specify prices for each event at which we are willing to sell or buy lottery tickets that pay $1 if the event occurs
  – Define Pr(A1), Pr(A2), Pr(A3) and Pr(A4) as the prices you are willing to sell and buy at
  – What are the properties of these prices so that I can NOT make you are sure loser?

  1. No negative prices: Pr(E) ≥ 0
  2. What is the price for all events? Pr(All) $= 1$
  3. How about the Union of A1 and A2? Pr(A1 $\cup$ A2) = Pr(A1) + Pr(A2)

- **Basic Theorems of Probability**: Using these three axioms, we can define all of the common theorems of probability.

  1. $\Pr(\emptyset) = 0$
  2. $\Pr(A^C) = 1 - \Pr(A)$
  3. For any event A, $0 \leq \Pr(A) \leq 1$.
  4. If $A \subset B$, then $\Pr(A) \leq \Pr(B)$.
  5. For any two events A and B, $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
  6. For any sequence of n events (which need not be disjoint) $A_1, A_2, \ldots, A_n$,
     $\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \Pr(A_i)$

- Examples: Let's assume we have an evenly-balanced, six-sided die. Then,
  1. Sample space $S = \{1, 2, 3, 4, 5, 6\}$
  2. $\Pr(1) = \cdots = \Pr(6) = 1/6$
  3. $\Pr(\emptyset) = \Pr(7) = 0$
  4. $\Pr(\{1, 3, 5\}) = 1/6 + 1/6 + 1/6 = 1/2$
  5. $\Pr\left(\overline{\{1, 2\}}\right) = \Pr(\{3, 4, 5, 6\}) = 2/3$
  6. Let $B = S$ and $A = \{1, 2, 3, 4, 5\} \subset B$. Then $\Pr(A) = 5/6 < \Pr(B) = 1$.
  7. Let $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$. Then $A \cup B = \{1, 2, 3, 4, 6\}$, $A \cap B = \{2\}$, and

$$\begin{aligned} \Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &= 3/6 + 3/6 - 1/6 \\ &= 5/6 \end{aligned}$$

1.4 Conditional Probability and Bayes' Law

- **Conditional Probability**: The conditional probability $\Pr(A|B)$ of an event A is the probability of A, given that another event B has occurred. It is calculated as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

- Example: Assume A and B occur with the following frequencies:

  and let $n_{ab} + n_{\bar{a}b} + n_{a\bar{b}} + n_{\overline{ab}} = N$. Then

	A	$\overline{A}$
B	n_{ab}	$n_{\bar{a}b}$
$\overline{B}$	$n_{a\bar{b}}$	$n_{\overline{ab}}$

  1. $\Pr(A) \approx \frac{n_{ab} + n_{a\bar{b}}}{N}$
  2. $\Pr(B) \approx \frac{n_{ab} + n_{\bar{a}b}}{N}$
  3. $\Pr(A \cap B) \approx \frac{n_{ab}}{N}$
  4. $\Pr(A|B) \approx \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{n_{ab}}{n_{ab} + n_{\bar{a}b}}$
  5. $\Pr(B|A) \approx \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{n_{ab}}{n_{ab} + n_{a\bar{b}}}$

- Example: A six-sided die is rolled. What is the probability of a 1, given the outcome is an odd number? Let $A = \{1\}$, $B = \{1, 3, 5\}$, and $A \cap B = \{1\}$. Then, $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/6}{1/2} = 1/3$.

- **Multiplicative Law of Probability**: The probability of the intersection of two events A and B is

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

  which follows directly from the definition of conditional probability.

- **Law of Total Probability**: Let S be the sample space of some experiment and let the disjoint k events $B_1, \ldots, B_k$ partition S. If A is some other event in S, then the events $AB_1, AB_2, \ldots, AB_k$ will form a partition of A and we can write A as

$$A = (AB_1) \cup \cdots \cup (AB_k)$$

Since the k events are disjoint,

$$\begin{aligned}\Pr(A) &= \sum_{i=1}^{k} \Pr(A, B_i) \\ &= \sum_{i=1}^{k} \Pr(B_i)\Pr(A|B_i)\end{aligned}$$

Sometimes it is easier to calculate the conditional probabilities and sum them than it is to calculate $\Pr(A)$ directly.

1.5 Bayesian thinking

- **Bayes Rule**: Assume that events $B_1, \ldots, B_k$ form a partition of the space S. Then

$$\Pr(B_j|A) = \frac{\Pr(A, B_j)}{\Pr(A)} = \frac{\Pr(B_j)\Pr(A|B_j)}{\sum\limits_{i=1}^{k} \Pr(B_i)\Pr(A|B_i)}$$

  If there are only two states of B, then this is just

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

- If this was a continuous distribution we could write this as:

$$\Pr(B_j|A) = \frac{\Pr(A, B_j)}{\Pr(A)} = \frac{\Pr(B)\Pr(A|B)}{\int\limits_{\infty}^{\infty} \Pr(A, B)\Pr(B)}$$

- Bayes rule determines the posterior probability of a state or type $\Pr(B_j|A)$ by calculating the probability $\Pr(AB_j)$ that both the event A and the state B_j will occur and dividing it by the probability that the event will occur regardless of the state (by summing across all B_i).

- Often Bayes' rule is used when one wants to calculate a posterior probability about the "state" or type of an object, given that some event has occurred. The states could be something like Normal/Defective, Normal/Diseased, Democrat/Republican, etc. The event on which one conditions could be something like a sampling from a batch of components, a test for a disease, or a question about a policy position.

- **Prior and Posterior Probabilities**: In the above, $\Pr(B_1)$ is often called the prior probability, since it's the probability of B_1 before anything else is known. $\Pr(B_1|A)$ is called the posterior probability, since it's the probability after other information is taken into account.

- Examples:

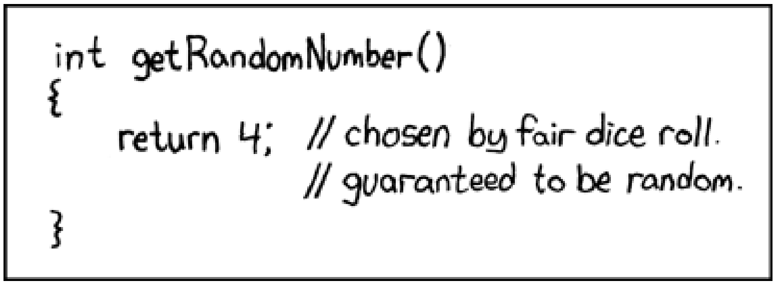
  1. A test for cancer correctly detects it 90% of the time, but incorrectly identifies a person as having cancer 10% of the time. If 10% of all people have cancer at any given time, what is the probability that a person who tests positive actually has cancer?

2. In Boston, 30% of the people are conservatives, 50% are liberals, and 20% are independents. In the last election, 65% of conservatives, 82% of liberals, and 50% of independents voted. If a person in Boston is selected at random and we learn that s/he did not vote last election, what is the probability s/he is a liberal?

1.6 Independence

- **Independence**: If the occurrence or nonoccurrence of either events A and B have no effect on the occurrence or nonoccurrence of the other, then A and B are independent. If A and B are independent, then
    1. $\Pr(A|B) = \Pr(A)$
    2. $\Pr(B|A) = \Pr(B)$
    3. $\Pr(A \cap B) = \Pr(A)\Pr(B)$
- **Pairwise independence**: A set of more than two events $A_1, A_2, \ldots, A_k$ is pairwise independent if $\Pr(A_i \cap A_j) = \Pr(A_i)\Pr(A_j)$, $\forall i \neq j$. Note that this does *not* necessarily imply that $\Pr(\bigcap_{i=1}^{k} A_i) = \prod_{i=1}^{K} \Pr(A_i)$.
- **Conditional independence**: If the occurrence of A or B conveys no information about the occurrence of the other, once you know the occurrence of a third event C, then A and B are conditionally independent (conditional on C):
    1. $\Pr(A|B \cap C) = \Pr(A|C)$
    2. $\Pr(B|A \cap C) = \Pr(B|C)$
    3. $\Pr(A \cap B|C) = \Pr(A|C)\Pr(B|C)$
- Conditional independence is one of the fundamental assumptions deployed for most statistical estimation techniques. It is a *very* strong assumption.

2 Random variables



```
int getRandomNumber()
{
    return 4;  // chosen by fair dice roll.
               // guaranteed to be random.
}
```

The intellectual beginnings of probability began in gambling, and this is still the easiest way to teach it. In probability theory, random variables are something abstract. A random variable is a yet-to-be observed value. What is the probability that a coin will turn up heads? What is the probability the next card will be an ace?

Depending on the kinds of events we are talking about, we have identified several "types" of random variables. These variables have known functional forms. Moreover, these functions have been extensively studied and their properties are well understood. The focus of this last section of lecturing is to get you familiar with these “kinds” of variables.

Don’t obsess about memorizing any of this. You will never be that far from Wikipedia. Focus on understanding:

- How this part of the lectures relates to the previous half of lectures
- Get a handle for the basic mapping of data types and the random variable “types” they go with (e.g., coin flips → binomial).
- The basic properties of random variables we care about (e.g., expected values, etc.)

2.1 Levels of Measurement

- In empirical research, data can be classified along several dimensions. We have already distinguished between discrete (countable) and continuous (uncountable) data. We can also look at the precision with which the underlying quantities are measured.
- **Nominal**: Discrete data are nominal if there is no way to put the categories represented by the data into a meaningful order. Typically, this kind of data represents names (hence ‘nominal’) or attributes, like Republican or Democrat.
- **Ordinal**: Discrete data are ordinal if there is a logical order to the categories represented by the data, but there is no common scale for differences between adjacent categories. Party identification is often measured as ordinal data.
- **Interval**: Discrete or continuous data are interval if there is an order to the values and there is a common scale, so that differences between two values have substantive meanings. Dates are an example of interval data.
- **Ratio**: Discrete or continuous data are ratio if the data have the characteristics of interval data and zero is a meaningful quantity. This allows us to consider the ratio of two values as well as difference between them. Quantities measured in dollars, such as per capita GDP, are ratio data.

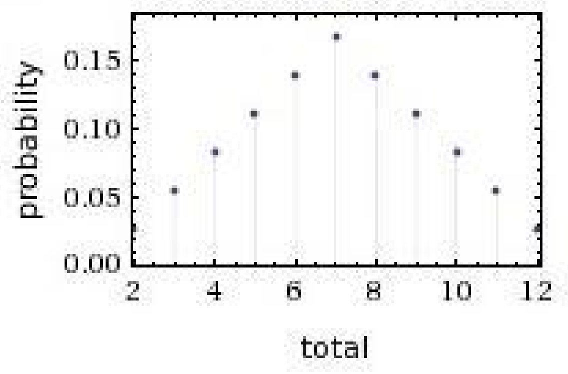
2.2 Discrete Distributions

- **Random Variable**: A random variable is a real-valued function defined on the sample space S; it assigns a real number to every outcome $s \in S$.
- **Discrete Random Variable**: Y is a discrete random variable if it can assume only a finite or countably infinite number of distinct values.
- Examples: number of wars per year, heads or tails, voting Republican or Democrat, number on a rolled die.
- **Probability Mass Function**: For a discrete random variable Y, the probability mass function (pmf)[‡] $p(y) = \Pr(Y = y)$ assigns probabilities to a countable number of distinct y values such that

[‡] Also referred to simply as the “probability distribution.”

1. $0 \leq p(y) \leq 1$
2. $\sum\limits_y p(y) = 1$

- **Example**: For one fair six-sided die, there is an equal probability of rolling any number. Since there are six sides, the probability mass function is then $p(y) = 1/6$ for $y = 1, \ldots, 6$. Each $p(y)$ is between 0 and 1. And, the sum of the $p(y)$'s is 1. If there are two six-sided dice, the probability mass function is shown below.



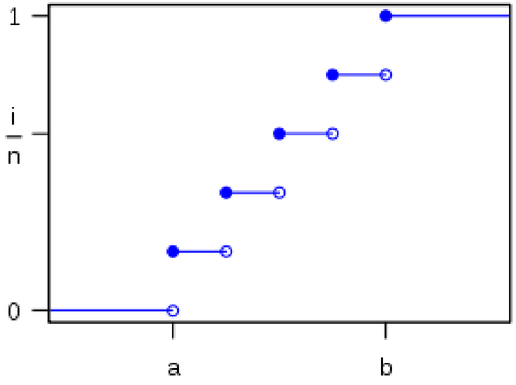
- **Cumulative Distribution**: The cumulative distribution $F(y)$ or $\Pr(Y \leq y)$ is the probability that Y is less than or equal to some value y, or

$$\Pr(Y \leq y) = \sum_{i \leq y} p(i)$$

. The CDF must satisfy these properties:

1. $F(y)$ is non-decreasing in y.
2. $\lim_{y \to -\infty} F(y) = 0$ and $\lim_{y \to \infty} F(y) = 1$
3. $F(y)$ is right-continuous.

- Example: For a fair die, $\Pr(Y \leq 1) = 1/6$, $\Pr(Y \leq 3) = 1/2$, and $\Pr(Y \leq 6) = 1$.



2.3 Continuous Distributions

- **Continuous Random Variable**: Y is a continuous random variable if there exists a non-negative function $f(y)$ defined for all real $y \in (-\infty, \infty)$, such that for any interval A,

$$\Pr(Y \in A) = \int_A f(y)dy$$

- Examples: age, income, GNP, temperature

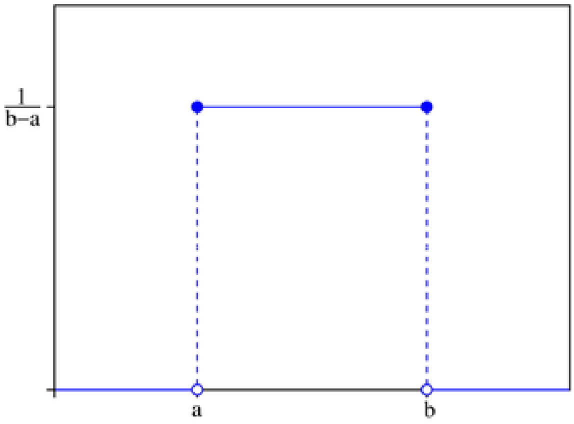
- **Probability Density Function**: The function f above is called the probability density function (pdf) of Y and must satisfy

  1. $f(y) \geq 0$
  2. $\int_{-\infty}^{\infty} f(y)dy = 1$

  Note also that $\Pr(Y = y) = 0$ — i.e., the probability of any point y is zero.

- **Example**: Uniform distribution (e.g., $f(x) = 1$).

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$



- **Cumulative Distribution**: Because the probability that a continuous random variable will assume any particular value is zero, we can only make statements about the probability of a continuous random variable being within an interval. The cumulative distribution gives the probability that Y lies on the interval $(-\infty, y)$ and is defined as

$$F(y) = \Pr(Y \leq y) = \int_{-\infty}^{y} f(s)ds$$

  Note that $F(y)$ has similar properties with continuous distributions as it does with discrete - non-decreasing, continuous (not just right-continuous), and $\lim_{y\to-\infty} F(y) = 0$ and $\lim_{y\to\infty} F(y) = 1$.

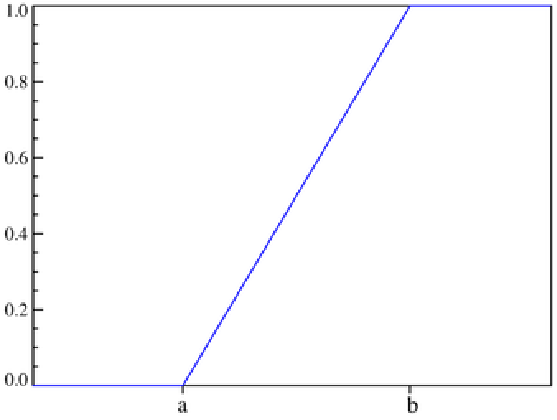
Similarly, we can also make probability statements about Y falling in an interval $a \leq y \leq b$.

$$\Pr(a \leq y \leq b) = \int\limits_a^b f(y)dy$$

- Example: $f(y) = 1, \quad 0 < y < 1$. Find $F(y)$ and $\Pr(.5 < y < .75)$.

$$F(y) = \int\limits_0^y f(s)ds = \int_0^y 1ds = s|_0^y = y$$

$$\Pr(.5 < y < .75) = \int_{.5}^{.75} 1ds = s|_{.5}^{.75} = .25$$



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

2.4 Joint Distributions

- Often, we are interested in two or more random variables defined on the same sample space. The distribution of these variables is called a joint distribution. Joint distributions can be made up of any combination of discrete and continuous random variables.

- Example: Suppose we are interested in the outcomes of flipping a coin and rolling a 6-sided die at the same time. The sample space for this process contains 12 elements:

$$\{h1, h2, h3, h4, h5, h6, t1, t2, t3, t4, t5, t6\}$$

  We can define two random variables X and Y such that $X = 1$ if heads and $X = 0$ if tails, while Y equals the number on the die. We can then make statements about the joint distribution of X and Y.

- **Joint discrete random variables**: If both X and Y are discrete, their joint probability mass function assigns probabilities to each pair of outcomes

$$p(x, y) = \Pr(X = x, Y = y)$$

Again, $p(x, y) \in [0, 1]$ and $\sum \sum p(x, y) = 1$.

If we are interested in the marginal probability of one of the two variables (ignoring information about the other variable), we can obtain the marginal pmf by summing across the variable that we don't care about:

$$p_X(x) = \sum_i p(x, y_i)$$

We can also calculate the conditional pmf for one variable, holding the other variable fixed. Recalling from the previous lecture that $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$, we can write the conditional pmf as

$$p_{Y|X}(y|x) = \frac{p(x, y)}{p_X(x)}, \quad p_X(x) > 0$$

- **Joint continuous random variables**: If both X and Y are continuous, their joint probability density function defines their distribution:

  $$\Pr((X, Y) \in A) = \iint_A f(x, y) dx dy$$

  Likewise, $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$.

  Instead of summing, we obtain the marginal probability density function by integrating out one of the variables:

  $$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

  Finally, we can write the conditional pdf as

  $$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}, \quad f_X(x) > 0$$

2.5 Expectation

- We often want to summarize some characteristics of the distribution of a random variable. The most important summary is the expectation (or expected value, or mean), in which the possible values of a random variable are weighted by their probabilities.

- **Expectation of Discrete Random Variable**: The expected value of a discrete random variable Y is

  $$E(Y) = \sum_y y p(y)$$

  In words, it is the weighted average of the possible values y can take on, weighted by the probability that y occurs. It is not necessarily the number we would expect Y to take on, but the average value of Y after a large number of repetitions of an experiment.

- Example: For a fair die,

  $$E(Y) = \sum_{y=1}^{6} y p(y) = \frac{1}{6} \sum_{y=1}^{6} y = 7/2$$

  We would never expect the result of a rolled die to be 7/2, but that would be the average over a large number of rolls of the die.

- **Expectation of a Continuous Random Variable**: The expected value of a continuous random variable is similar in concept to that of the discrete random variable, except that instead of summing using probabilities as weights, we integrate using the density to weight. Hence, the expected value of the continuous variable Y is defined by

$$E(Y) = \int_{-\infty}^{\infty} y f(y) dy$$

- Example: Find $E(Y)$ for $f(y) = \frac{1}{1.5}, \quad 0 < y < 1.5$.

$$E(Y) = \int_{0}^{1.5} \frac{1}{1.5} y dy = \frac{1}{3} y^2 \Big|_{0}^{1.5} = .75$$

- **Expected Value of a Function**:

  1. Discrete: $E[g(Y)] = \sum_y g(y)p(y)$
  2. Continuous: $E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy$

- **Other Properties of Expected Values**:

  1. $E(c) = c$
  2. $E[E[Y]] = E[Y]$ (because the expected value of a random variable is a constant)
  3. $E[cg(Y)] = cE[g(Y)]$
  4. $E[g(Y_1) + \cdots + g(Y_n)] = E[g(Y_1)] + \cdots + E[g(Y_n)]$

- **Variance**: We can also look at other summaries of the distribution, which build on the idea of taking expectations. Variance tells us about the "spread" of the distribution; it is the expected value of the squared deviations from the mean of the distribution. The standard deviation is simply the square root of the variance.

  1. Variance: $\sigma^2 = \text{Var}(Y) = E[(Y - E(Y))^2] = E(Y^2) - [E(Y)]^2$
  2. Standard Deviation: $\sigma = \sqrt{\text{Var}(Y)}$

- **Covariance and Correlation**: The covariance measures the degree to which two random variables vary together; if the covariance is positive, X tends to be larger than its mean when Y is larger than its mean. The covariance of a variable with itself is the variance of that variable.

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

  The correlation coefficient is the covariance divided by the standard deviations of X and Y. It is a unitless measure and always takes on values in the interval $[-1, 1]$.

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}$$

- **Conditional Expectation**: With joint distributions, we are often interested in the expected value of a variable Y if we could hold the other variable X fixed. This is the conditional expectation of Y given $X = x$:

  1. Y discrete: $E(Y|X = x) = \sum_y y p_{Y|X}(y|x)$
  2. Y continuous: $E(Y|X = x) = \int_y y f_{Y|X}(y|x) dy$

  The conditional expectation is often used for prediction when one knows the value of X but not Y; the realized value of X contains information about the unknown Y so long as $E(Y|X = x) \neq E(Y) \forall x$.

2.6 Special Discrete Distributions

- **Binomial Distribution**: Y is distributed binomial if it represents the number of "successes" observed in n independent, identical "trials," where the probability of success in any trial is p and the probability of failure is $q = 1 - p$.

  For any particular sequence of y successes and $n - y$ failures, the probability of obtaining that sequence is $p^y q^{n-y}$ (by the multiplicative law and independence). However, there are $\binom{n}{y} = \frac{n!}{(n-y)!y!}$ ways of obtaining a sequence with y successes and $n - y$ failures. So the binomial distribution is given by

  $$p(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, 2, \ldots, n$$

  with mean $\mu = E(Y) = np$ and variance $\sigma^2 = V(Y) = npq$.

- Example: Republicans vote for Democrat-sponsored bills 2% of the time. What is the probability that out of 10 Republicans questioned, half voted for a particular Democrat-sponsored bill? What is the mean number of Republicans voting for Democrat-sponsored bills? The variance?

  1. $p(5) = \binom{10}{5}(.02)^5(.98)^5 = .073$
  2. $E(Y) = np = 10(.02) = .2$
  3. $V(Y) = npq = 10(.02)(.98) = .196$

- **Poisson Distribution**: A random variable Y has a Poisson distribution if

  $$p(y) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y = 0, 1, 2, \ldots, \quad \lambda > 0$$

  The Poisson has the unusual feature that its expectation equals its variance: $E(Y) = V(Y) = \lambda$. The Poisson distribution is often used to model event counts: counts of the number of events that occur during some unit of time. λ is often called the "arrival rate."

- Example: Border disputes occur between two countries at a rate of 2 per month. What is the probability of 0, 2, and less than 5 disputes occurring in a month?

  1. $p(0) = \frac{2^0}{0!} e^{-2} = .13$
  2. $p(2) = \frac{2^2}{2!} e^{-2} = .27$
  3. $\Pr(Y < 5) = \sum\limits_{y=0}^{4} \frac{2^y}{y!} e^{-2} = .95$

2.7 Special Continuous Distributions

- **Uniform Distribution**: A random variable Y has a continuous uniform distribution on the interval (α, β) if its density is given by

$$f(y) = \frac{1}{\beta - \alpha}, \quad \alpha \leq y \leq \beta$$

  The mean and variance of Y are $E(Y) = \frac{\alpha+\beta}{2}$ and $V(Y) = \frac{(\beta-\alpha)^2}{12}$.

- Example: Y uniformly distributed over $(1, 3)$.

- **Normal Distribution**: A random variable Y is normally distributed with mean $E(Y) = \mu$ and variance $V(Y) = \sigma^2$ if its density is

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

- Example: Y normally distributed with mean $\mu = 0$ and variance $\sigma^2 = .1$

2.8 Summarizing Observed Data

- So far, we've talked about distributions in a theoretical sense, looking at different properties of random variables. We don't observe random variables; we observe realizations of the random variable.

- **Central tendency**: The central tendency describes the location of the "middle" of the observed data along some scale. There are several measures of central tendency.

  1. **Sample mean**: This is the most common measure of central tendency, calculated by summing across the observations and dividing by the number of observations.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

     The sample mean is an estimate of the expected value of a distribution.

  2. **Sample median**: The median is the value of the "middle" observation. It is obtained by ordering n data points from smallest to largest and taking the value of the $n + 1/2$th observation (if n is odd) or the mean of the $n/2$th and $(n/2) + 1$th observations (if n is even).

  3. **Sample mode**: The mode is the most frequently observed value in the data:

$$m_x = X_i : n(X_i) > n(X_j) \forall j \neq i$$

     When the data are realizations of a continuous random variable, it often makes sense to group the data into bins, either by rounding or some other process, in order to get a reasonable estimate of the mode.

  4. Exercise: Calculate the sample mean, median, and mode for the following two variables, X and Y.

X	6	3	7	5	5	5	6	4	7	2
Y	1	2	1	2	2	1	2	0	2	0

- **Dispersion**: We also typically want to know how spread out the data are relative to the center of the observed distribution. Again, there are several ways to measure dispersion.

  1. **Sample variance**: The sample variance is the sum of the squared deviations from the sample mean, divided by the number of observations minus 1.

     $$\text{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

     Again, this is an estimate of the variance of a random variable; we divide by $n-1$ instead of n in order to get an unbiased estimate.

  2. **Standard deviation**: The sample standard deviation is the square root of the sample variance.

     $$\text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  3. **Median absolute deviation (MAD)**: The MAD is a different measure of dispersion, based on deviations from the median rather than deviations from the mean.

     $$MAD(X) = median(|x_i - median(x)|)$$

  4. Exercise: Calculate the sample variance, standard deviation, and MAD for the following two variables, X and Y.

X	6	3	7	5	5	5	6	4	7	2
Y	1	2	1	2	2	1	2	0	2	0

- **Covariance and Correlation**: Both of these quantities measure the degree to which two variables vary together, and are estimates of the covariance and correlation of two random variables as defined above.

  1. **Sample covariance**: $\text{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
  2. **Sample correlation**: $r = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$
  3. Exercise: Calculate the sample covariance and correlation coefficient for the following two variables, X and Y.

X	6	3	7	5	5	5	6	4	7	2
Y	1	2	1	2	2	1	2	0	2	0