Math Camp 2017 - Probability*

James Banovetz

Department of Economics, UC Santa Barbara

September 11, 2017

1. **Sets**

   (a) Definition. Let S be the sample space, i.e., the set of all elements under consideration. Let A and B be sets contained in S. Then:

   i. If every point in A is also in B, then A is a **subset** of B, denoted $A \subset B$

   ii. The **empty set** contains no points, denoted $\{\emptyset\}$

   iii. The **union** of A and B is the set of points in A, B, or both, denoted $A \cup B$

   iv. The **intersection** of A and B is the set of points in both A and B, denoted $A \cap B$

   v. The **complement** of A is the set of elements in S but not in A, denoted A^c

   vi. If A and B have no points in common, they are **disjoint** if $A \cap B = \emptyset$

   (b) Theorem (CB THM 1.1.4). For any sets A, B, and C defined on the sample space S:

   - Commutativity: $A \cup B = B \cup A$, $A \cap B = B \cap A$
   - Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$
   - Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
   - DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$

   (c) Definition. Let S be a sample space. A collection of subsets of S is called a **sigma algebra** (or Borel field), denoted $\mathcal{B}$, if it satisfies three properties:

   i. $\emptyset \in \mathcal{B}$

   ii. If $A \in \mathcal{B}$ then $A^c \in \mathcal{B}$

   iii. If $A_1, A_2, \cdots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$

   (d) Example. Consider the sample space $S = \{1, 2, 3\}$. One sigma algebra is known as the trivial sigma algebra, given by $\{\emptyset, S\}$. The one we'll typically be concerned with is $\mathcal{B} =$

*These lecture notes are drawn principally from *Statistical Inference*, 2nd ed., by George Casella and Roger L. Berger.

{all subsets of S}, i.e., the power set of S. In this case, there are $n = 3$ elements, so there are $2^3 = 8$ subsets, the collection of which forms the sigma algebra $\mathcal{B}$:

$$\begin{array}{lll} \{1\} & \{1,2\} & \{1,2,3\} \\ \{2\} & \{1,3\} & \emptyset \\ \{3\} & \{2,3\} & \end{array}$$

This ties into what follows, as we will be concerned with assigning probabilities to every set in the power set (e.g., what's the probability of 1, of 2, of 1 and 2, etc.).

2. **Probabilities**

(a) Definition. Given a sample space S and an associated sigma algebra $\mathcal{B}$, a **probability function** is a function $\mathbb{P}$ with domain $\mathcal{B}$ that maps to [0,1], i.e. $\mathbb{P} : \mathcal{B} \to [0, 1]$, that satisfies

i. $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{B}$

ii. $\mathbb{P}(S) = 1$

iii. If $A_1, \ldots A_n \in \mathcal{B}$ are pairwise disjoint, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

These are known as the **axioms of probability**

(b) Example. Suppose we have a fair coin. Then the sample space is $S = \{H, T\}$ (i.e., heads or tails). If we define heads and tails to each have a probability of one-half, then:

i. $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2} \geq 0$

ii. $\mathbb{P}(S) = 1$ (i.e., the probability we get heads, tails, both, or neither, is equal to 1)

iii. $\mathbb{P}(H \cup T) = \frac{1}{2} + \frac{1}{2}$

(c) Theorem: If P is a probability function and A and B is any set in $\mathcal{B}$, then

i. $\mathbb{P}(\emptyset) = 0$

ii. $\mathbb{P}(A) \leq 1$

iii. $\mathbb{P}(A^c) = 1 - P(A)$

iv. $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$

v. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

vi. If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$

3. **Counting**

(a) Aside. A topic intimately related to the basics of probability theory is the idea of counting. When trying to calculate something like $\mathbb{P}(A)$, we can theoretically follow simple steps:

i. List each element in our set S

ii. Assign probabilities to elements in S

iii. Define A to be a set of elements in S

iv. Sum the probabilities in each event in A

This is easy to do with something like coin-flipping, but in practice can be vastly more difficult. We'll cover four basic scenarios and the associated formulas.

(b) Theorem. If there are k groups with the ith group containing n_i elements for groups $i = 1, \ldots, k$, then there are $n_1 \times n_2 \times \cdots \times n_k$ ways to form k-tuples continaing one element from each group. This is known as the fundamental theorem of counting.

(c) Example. Suppose that license plates are created using three letters (A-Z) followed by four numerical digits (0-9). If repeated letters/digits are allowed, how many distinct license plates are there?

$$26 \times 26 \times 26 \times 10 \times 10 \times 10 \times 10 \approx 175 \text{ million}$$

(d) Definition. The **factorial** of a natural number $n \in \mathbb{N}$ is the production of all natural numbers less than or equal to n, that is,

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 = \prod_{i=1}^{n} i$$

(e) Aside. There are four canonical ways of counting the total number of possibilities N when there are n items from which to choose and we are choosing r times.

i. **Ordered, Without Replacement** (also known as a **permutation**)

$$N = P_r^n = \frac{n!}{(n-r)!}$$

ii. Example. Padlock "Combinations". Simple padlocks feature 40 digits, requiring three distinct digits in the proper order to unlock. How many possible padlock "combinations" are there?

$$N = 40 \times 39 \times 38 = \frac{40!}{37!} = 59,280$$

iii. **Ordered, With Replacement.** This corresponds the fundamental theorem of counting, where $n_i = n_j$ for all i and j.

$$N = n^r$$

iv. Example. Recall our license plates example from before. Some states give trucks only numerical digits, where duplicates are allowed but order matters. If a license plate has six numerical digits (0-9), how many different truck license plates are there?

$$N = 10^6 = 1,000,000$$

v. **Unordered, Without Replacement** (also known as a **combination**)

$$N = C_r^n = \binom{n}{r} = \frac{n!}{(n-r)!r!}$$

vi. Example. Suppose you have 5 positions in your PhD program, but 30 applicants. How many different incoming classes could you select?

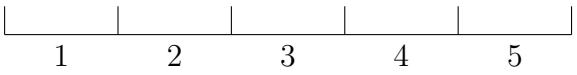
$$N = \binom{30}{5} = \frac{30!}{(25!)(5!)} = 142,506$$

vii. **Unordered, With Replacement**

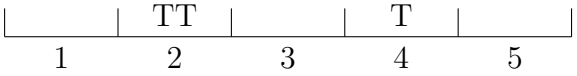
$$N = \frac{(n+r-1)!}{(n-1)!r!} = \binom{n+r-1}{r}$$

viii. Example. Suppose we have five potential job sites, enumerate 1-5, and three identical trucks (in the sense that it does not matter which truck goes to which site, what matters is the number of trucks that end up at a site). If multiple trucks can be sent to the same site, how many different assignments are possible?

- Think about the 5 sites as "bins," numbered 1 through 5

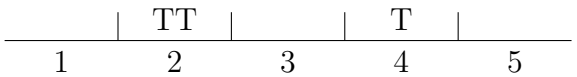


- Imagine trucks can go to different sites, e.g.:



This would correspond to two trucks at site 2 and one at site 4 (alternatively, this could be thought of as the outcome where two 2's are drawn and one 4 is drawn).

- Think of each bin "wall" and each truck as an element to be ordered. Note that the first and last walls are "immobile", so we'll forget them:



This corresponds to the ordering $WTTWWTW$.

- Now we have seven total positions. If they were distinct elements, we'd have 7! possibilities. Walls and Trucks are indistinguishable from other walls and trucks, respectively, so we need to divide out the redundancies:

$$N = \frac{7!}{4!3!} = \binom{7}{3}$$

which corresponds to our formula for unordered, with replacement, when we have five objects, picking three!

These tools are helpful when a sample space S is finite and all outcomes are equally likely. If there are n elements in $S = \{s_1, \ldots, s_n\}$ and $P(s_i) = 1/N$, then for a set of outcomes A:

$$\mathbb{P}(A) = \frac{\text{\# of elements in } A}{\text{\# of elements in } S}$$

This idea was and is a building block of probability theory.

(f) Example. Consider a typical deck of cards: 52 total cards, with 4 suits (clubs, diamonds, hearts, spades) of 13 cards each. If you draw five cards, what is the probability of drawing a full house (one pair and one three-of-a-kind, e.g., 2 kings and 3 aces)?

- First step: pick a type of card from the 13 denominations: $\binom{13}{1}$
- Second step: pick two of the four suits: $\binom{4}{2}$
- Third step: pick another typ from the remaining 12 denominations: $\binom{12}{1}$
- Fourth step: pick thee of the four suits: $\binom{4}{3}$

The total number of ways to pick a pair and a three-of-a-kind is the product of these four combinations. The total number of ways to draw five cards is $\binom{52}{5}$. Thus, the probability is:

$$\mathbb{P}(\text{full house}) = \frac{\binom{13}{1}\binom{4}{2}\binom{12}{1}\binom{4}{3}}{\binom{52}{5}} \approx 0.00144$$

4. **Conditional Probabilities and Independence**.

(a) Aside. Once we've established the basics of probability theory, we can start thinking about how to update and fold new information into the probabilities. To formalize th notion of updating via new information, we think about conditional probabilities.

(b) Definition. If A and B are events in S, and $\mathbb{P}(B) > 0$, then the **conditional probability** of A given B, denoted $\mathbb{P}(A|B)$, is defined as

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

(c) Example. Suppose we toss a fair six-sided die once. What is the probability that we observe a 1, given that we observe an odd number?

$$\mathbb{P}(\text{odd}) = 1/2 \qquad \text{(three odds out of six)}$$

$$\mathbb{P}(1 \text{ and an odd}) = 1/6 \qquad \text{(one 1 out of six total)}$$

$$\mathbb{P}(\text{one}|\text{odd}) = \frac{\mathbb{P}(\text{one and an odd})}{\mathbb{P}(\text{odd})} \qquad \text{(by def. of cond. prob.)}$$

$$\mathbb{P}(\text{one}|\text{odd}) = \frac{1/6}{1/2} = 1/3$$

(d) Definition. Two events A and B in S are said to be **independent** if and only if we have one of three equivalent conditions:

- $\mathbb{P}(A|B) = \mathbb{P}(A)$
- $\mathbb{P}(B|A) = \mathbb{P}(B)$
- $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

(e) Theorem (CB THM 1.3.5). Let A and B be events in a sample space S. Then the following relationship holds between conditional probabilities:

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

This is known as Bayes' Rule.

(f) Theorem. Let A be an event in our probability space, and let $\{B_1, \ldots, B_n\}$ be pairwise disjoint events whose union is the entire sample space. Then

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

(g) Example. Consider testing for the presence of a disease. The test is very accurate in the sense that if a patient has the disease, the test always comes back positive, i.e., $\mathbb{P}(\text{positive}|\text{disease}) = 1$. Sometimes the test is inaccurate, however, in the sense that the test gives a false positive (a postive value when a person doesn't have the disease) with probability 0.005, i.e., $\mathbb{P}(\text{positive}|\text{no disease}) = 0.005$. If the probability of having the disease is 0.001, i.e., $\mathbb{P}(\text{disease}) = 0.001$, what is the probability a patient has the disease, given they have a positive test?

By our law of total probability, we know that:

$$\begin{aligned}\mathbb{P}(\text{positive}) &= \mathbb{P}(\text{disease})\mathbb{P}(\text{positive}|\text{disease}) + \mathbb{P}(\text{no disease})\mathbb{P}(\text{positive}|\text{no disease}) \\ &= 0.001 * 1 + 0.999 * 0.005 = 0.005995\end{aligned}$$

Then by Bayes' rule, we have that:

$$\begin{aligned}\mathbb{P}(\text{disease}|\text{positive}) &= \mathbb{P}(\text{positive}|\text{disease})\frac{\mathbb{P}(\text{disease})}{\mathbb{P}(\text{positive})} \\ &= 1 * \frac{0.001}{0.005995} \approx 0.1668\end{aligned}$$

(h) Aside. This problem, posed to people untrained in probability theory, generates inaccurate answers (usually in the neighborhood of 95% rather than a guess in the area of 1/6). In behavioral economics and psycology, this is known as base rate neglect or the base rate fallacy.

5. **Random Variables and Distribution Functions**

(a) Definition. A **random variable** is a function that maps a sample space S onto the real numbers, e.g., $R : S \to \mathbb{R}$.

(b) Example. For die rolls, we can define the set $\{1, 2, 3, 4, 5, 6\}$ as the events in the sample space, then define the random variable X as

$$X = \begin{cases} 1 & \text{if we observe an even value} \\ 0 & \text{if we observe an odd value} \end{cases}$$

(c) Aside. Further, we can define a probability function over the random variables. If $X = x_i$ if and only if we observe an outcome in $s_j \in S$ such that $X(s_j) = x_i$, then

$$P_X(X = x_i) = P(\{s_j \in S | X(s_j) = x_i\})$$

From our example above, $P(X = 1) = 1/2$, where our set of s_j are $\{2, 4, 6\}$. Note that we need to be careful with our notation—for unrealized values of a random variable, an uppercase is used (e.g., X). For realized outcomes, we use lower case (e.g., x).

(d) Definition. The **cumulative distribution function** or **cdf** of a random variable X, denoted $F_X(x)$, is defined as

$$F_X(x) = P_X(X \leq x), \quad \text{for all } x \in \mathbb{R}$$

(e) Example. Consider the experiment where we're tossing a coin twice, and our RV is $X =$ the number of heads. Then the cdf of X is

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \frac{1}{4} & \text{if } 0 \leq x < 1 \\ \frac{3}{4} & \text{if } 1 \leq x < 2 \\ 1 & \text{if } 2 \leq x < \infty \end{cases}$$

(f) Theorem (CB THM 1.5.3). The function $G(x)$ is a CDF if and only if it satisfies three conditions:

i. $\lim_{x\to-\infty} G(x) = 0$ and $\lim_{x\to\infty} G(x) = 1$

ii. $G(x)$ is a nondecreasing function of x

iii. $G(x)$ is right-continuous (i.e., for every number x_0, $\lim_{x\downarrow x_0} G(x) = G(x_0)$

(g) Definition. The random variables X and Y are **identically distributed** if, for every set $A \in \mathcal{B}$, $P(X \in A) = P(Y \in A)$. In other words:

$$X \text{ and } Y \text{ are identically distributed } \iff F_X(x) = F_y(x) \quad \forall\, x$$

(h) Definition. A random variable X is **continuous** if $F_X(x)$ is a continuous function of x. A random variable is **discrete** if $F_X(x)$ is a step function of x.

(i) Example. Consider a simple CDF for a continuous random variable (this is from an exponential distribution):

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x} & \text{if } x \geq 0 \end{cases}$$

Similarly, consider the CDF for a discrete Bernoulli random variable (a single 0 or 1, where 1 occurs with probability p):

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x < \infty \end{cases}$$

(j) Definition. The **probability mass function** (or pmf) of a discrete random variable X is given by $f_X(x) = P(X = x)$ for all x.

(k) Example. Suppose you're betting on who wins opening week this NFL season. If you're guessing based on a total lack of knowledge, the probability you're correct on any given game is 1/2. If there are 16 games, what's the probability you'll get x games correct?

You have sixteen guesses, each with a probability of 1/2 being correct: if you get x correct, each with probability 1/2, that will give something like:

$$\left(\frac{1}{2}\right)^x$$

Further, you get $16 - x$ incorrect, each with probability $(1 - 1/2)$. Combined with the above, we can think about something like:

$$\left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{16-x}$$

However, there are "16 choose x" ways for you to get your x guesses correct. Thus:

$$P(X = x) = \binom{16}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{16-x}$$

It turns out that this is an example of the binomial distribution. with $n = 16$ and $p = 1/2$.

(l) Definition. The **probability density function** (or pdf) of a continuous random variable X is given by $f_X(x)$, where

$$\int_{-\infty}^{x} f_X(t)dt = F_X(x) \quad \forall\, x$$

Further, note that if $f_X(x)$ is continuous, then $\frac{d}{dx}F_X(x) = f_X(x)$ by the fundamental theorem of calculus.

(m) Example. The PDF for a uniform $[0, 1]$ variables is:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{if } x > 1 \end{cases} \qquad \text{or} \qquad f_X(x) = 1 \cdot \mathbb{1}\{0 \le x \le 1\}$$

(n) Theorem. A function $f_X(x)$ is a pdf or pmf of a random variable X if and only if

i. $f_X(x) \ge 0$ for all x

ii. $\sum_x f_X(x) = 1$ or $\int_{-\infty}^{\infty} f_X(x)dx = 1$

(o) Aside. The **support** is the subset of the domain of $f_X(x)$ where the function is strictly positive. Elsewhere, $f_X(x)$ takes on a value of zero. In the previous example, the support is $[0, 1]$. For the standard normal, the support is $(-\infty, \infty)$. *Always* remember to include the support when writing down PDFs, as it becomes extremely important when calculating moments, transforming variables, etc. etc.

6. **Moments**

(a) Definition. The **expected value** of a random variable $g(X)$, denoted by $\mathbb{E}\big[g(X)\big]$, is given given by:

$$\mathbb{E}\big[g(X)\big] = \begin{cases} \int_{-\infty}^{\infty} g(x)f_X(x)dx & \text{if } X \text{ is continuous} \\ \sum_x g(x)f_X(x) & \text{if } X \text{ is discrete} \end{cases}$$

so long as $\mathbb{E}\Big[|g(X)|\Big] < \infty$.

(b) Example. Find the expected value of X, where X is distributed exponentially (β), i.e., $f_X(x) = \beta e^{-\beta x}$, $0 \le x < \infty$.

$$\begin{aligned} \mathbb{E}\big[X\big] &= \int_0^{\infty} x\beta e^{-\beta x}dx && \text{(by def. of } \mathbb{E}) \\ &= \big[x\left(-e^{-\beta x}\right)\big]_0^{\infty} + \int_0^{\infty} e^{-\beta x}dx && \text{(by integration by parts)} \\ &= 0 + \int_0^{\infty} e^{-\beta x}dx && (e^{-\beta x} \to 0 \text{ faster than } x \text{ grows}) \\ &= \left[-\frac{1}{\beta}e^{-\beta x}\right]_0^{\infty} && \text{(taking the integral)} \\ \mathbb{E}\big[X\big] &= \frac{1}{\beta} && \text{(evaluating)} \end{aligned}$$

(c) Theorem (CB THM 2.2.5). Expectations are linear, i.e., for a random variable X and any constants a, b, and c,

$$\mathbb{E}\big[ag(X) + bh(X) + c\big] = a\mathbb{E}\big[g(X)\big] + b\mathbb{E}\big[h(X)\big] + c$$

(d) Definition. For each integer n, the nth **moment** of X, m_n, is

$$m_n = \mathbb{E}\big[X^n\big]$$

The nth **central moment**, μ_n, is

$$\mu_n = \mathbb{E}\big[(X-\mu)^n\big]$$

Where $m_1 = \mu_1 = \mathbb{E}\big[X\big]$.

(e) Definition. The **variance** of a random variable X is defined to be the expectation:

$$\text{Var}\big[X\big] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$$

This can equivalently be written as $\text{Var}\big[X\big] = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2$.

(f) Theorem (CB THM 2.3.4). If X is a random variable with finite variance, then for any constants a and b,

$$\text{Var}[aX+b] = a^2\text{Var}[X]$$

(g) Aside. There are other moments, e.g., skewness (a third) and kurtosis (a fourth), but we rarely deal with them. Virtually never during the first year of a PhD Economics degree. That said, we are very interested in the function that will produce higher-order moments.

(h) Definition. Let X be a random variable with CDF $F_X(x)$. The **moment generating function** or MGF of X, denoted by $M_X(t)$, is given by

$$M_X(t) = \mathbb{E}\big[e^{tx}\big]$$

if the expectation exists for t in the neighborhood of 0.

(i) Example. Find the MGF for a random variable X with PDF $f(x) = \beta e^{-\beta x}$, $x \in [0,\infty)$:

$$\begin{aligned}
\mathbb{E}\big[e^{tX}\big] &= \int_0^\infty e^{tx}\cdot\beta e^{-\beta x}dx && \text{(finding the expected value)}\\
&= \int_0^\infty \beta e^{-(\beta-t)x}dx && \text{(simplifying)}\\
&= \left[-\frac{\beta}{\beta-t}e^{-(\beta-t)x}\right]_0^\infty && \text{(integrating)}\\
&= \big[0\big] - \left[-\frac{\beta}{\beta-t}e^{-(\beta-t)0}\right] && \text{(evaluating)}\\
M_X(t) &= \frac{\beta}{\beta-t} && \text{(simplifying)}
\end{aligned}$$

Note that $t < \beta$ for these integrals to be finite, which works via our assumption that t is "very close" to 0.

(j) Aside. We call these moment generating functions becasue of the following property:

$$\mathbb{E}\big[X^n\big] = M_X^{(n)}(0) = \frac{d^n}{dt^n}M_X(t)\Big|_{t=0}$$

Assuming that we can exchange integrals and derivatives (which we can, almost always, in our classes during the first year), we can show that this is true for the expected value:

$$\begin{aligned}
\frac{d}{dt}M_X(t) &= \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f_X(x)dx \\
&= \int_{-\infty}^{\infty} \left(\frac{d}{dt}e^{tx}\right) f_X(x)dx \\
&= \int_{-\infty}^{\infty} xe^{tx} f_X(x)dx \\
&= \mathbb{E}\big[Xe^{tX}\big] \\
\frac{d}{dt}M_X(t)\Big|_{t=0} &= \mathbb{E}\big[Xe^{tX}\big]\Big|_{t=0} \\
&= \mathbb{E}\big[X\big]
\end{aligned}$$

Proceeding via induction, we could prove that this holds for any integer n, supposing that the MGF exists. In other words, we could use MGFs to give us any/all non-central moments m_n.

(k) Example. Consider the MGF of the exponential RV we found above:

$$\begin{aligned}
M_X(t) &= \beta[\beta - t]^{-1} \\
M_X^{(1)}(t) &= -\left(-\beta[\beta - t]^{-2}\beta\right) && \text{(differentiating w.r.t. } t\text{)} \\
&= \frac{\beta}{(\beta - t)^2} && \text{(simplifying)} \\
M_X^{(1)}(0) &= \frac{\beta}{\beta^2} && \text{(evaluating at } t = 0\text{)} \\
&= \frac{1}{\beta} && \text{(simplifying)}
\end{aligned}$$

Which is the expected value we found before.

(l) Theorem (CB THM 2.3.11 & 2.3.12). For a random variable X, it is the case that

- If $M_X(t)$ exists for a distribution $F_X(t)$, it is unique
- If a sequence of MGFs exists, convergence in MGFs implies convergence in CDFs

(m) Theorem(CB THM 2.3.15). Let $Y_1, \ldots, Y_n$ be independent random variables with MGFs $M_{Y_1}(t), \ldots, M_{Y_n}(t)$. Then if $U = k_1Y_1 + k_2Y_2 + \cdots + k_nY_n + \text{c}$, then the MGF of U is

$$M_U(t) = e^{ct}\prod_{i=1}^{n} M_{Y_i}(k_i\, t)$$

(n) Aside. Note that we haven't yet defined the concept of independence. For now, it is enough to assert that the variables Y_i and Y_j contain no relevent information about each other.

(o) Example. Using MGFs, show that the sum of two normal random variables is also a normal random variable. Suppose we have $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$. Then the MGFs are

$$M_X(t) = \exp\left\{\mu_x t + \frac{1}{2}\sigma_x^2 t^2\right\} \qquad \text{and} \qquad M_Y(t) = \exp\left\{\mu_y t + \frac{1}{2}\sigma_y^2 t\right\}$$

Then if $Z = X + Y$, the MGF of Z is $M_Z(t) = M_X(t)M_Y(t)$:

$$\begin{aligned} M_Z(t) &= \exp\left\{\mu_x t + \frac{1}{2}\sigma_x^2 t^2\right\} \exp\left\{\mu_y t + \frac{1}{2}\sigma_y^2 t\right\} \\ &= \exp\left\{(\mu_x + \mu_y)t + \frac{1}{2}(\sigma_x^2 + \sigma_y^2)t^2\right\} \end{aligned}$$

Since we know that MGFs uniquely identify distributions, this implies that the sum of normal RVs is also normal, i.e., $Z \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

7. **Transformations**.

(a) Aside. We're frequently more interested in the distribution of functions of random variables than in the parent distributions themselves. If X is a random variable, we often want to go about finding the distribution of $g(X)$. This leads us to the concept of transformations.

(b) Definition. Let X be a random variable with CDF $F_X(x)$. Then a function of X, $Y = g(X)$ is also a random variable, known as the **transformation** of X. Moreover, For any set A,

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) = \mathbb{P}(X \in g^{-1}(A))$$

which defines the probability distibution of $Y = g(X)$.

(c) Example. Let X be a discrete random variable following a binomial distribution, i.e.,

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

where n is a positive integer and $p \in [0, 1]$. Consider the random variable $Y = g(X)$, where $g(X) = n - X$. We can rearrange to get $X = n - Y$. Using the defintion above, we can find the PMF of Y:

$$\begin{aligned} f_Y(y) &= \mathbb{P}(Y = y) && (Y \text{ is discrete}) \\ &= \mathbb{P}(n - X = y) && (\text{by def. of } Y) \\ &= \mathbb{P}(X = n - y) && (\text{rearranging}) \\ &= f_X(n - y) && (\text{by def. of the PMF}) \\ &= \binom{n}{y - n} p^{n-y}(1-p)^{n-(n-y)} && (\text{plugging in values}) \\ f_Y(y) &= \binom{n}{y} (1-p)^y p^{n-y}, \quad y = 0, 1, \ldots, n && (\text{simplifying}) \end{aligned}$$

Note the switch in the combination; recall our counting definitions for a justification. Thus, the transformation of X also has a binomial distribution.

(d) Aside. While this can be a straightforward excercise for discrete random variables (although it may not always be this easy), we will spend more time dealing with transformations of continuous random variables during the first year. For univariate transformations, we can follow a few simple steps to get our transformation, using the definition of the transformation:

i. Let U be a function of Y, i.e., $U = g(Y)$

ii. Consider the probability that $U \leq u$

iii. Substitute in $g(Y)$ for U and isolate Y (pay attention to supports)

iv. Go from probabilities to CDFs

v. Differentiated w.r.t. u to find $f_U(u)$

(e) Example. Consider a random variable Y with CDF $F_Y(y)$ and support $(-\infty, \infty)$. We can find and expression for $f_U(u)$, where $U = Y^2$:

$$\begin{aligned} P(U \leq u) &= P(Y^2 \leq u) && \text{(plugging in for } U\text{)} \\ &= P(-\sqrt{u} \leq Y \leq \sqrt{u}) && \text{(isolating } Y\text{)} \\ &= F_Y(\sqrt{u}) - F_Y(-\sqrt{u}) && \text{(by our properties of CDFs)} \\ f_U(u) &= \left(\frac{1}{2\sqrt{u}}\right) f_Y(\sqrt{u}) + \left(\frac{1}{2\sqrt{u}}\right) f_Y(-\sqrt{u}) && \text{(differentiating w.r.t. } u\text{)} \\ &= \left(\frac{1}{2\sqrt{u}}\right) \big[f_Y(\sqrt{u}) + f_Y(-\sqrt{u})\big], \quad u \in [0, \infty) && \text{(simplifying)} \end{aligned}$$

Note that this can get more complicated if the support is not symmetric around zero.

(f) Example. Consider a random variable X with CDF $F_X(x)$ and support (-2,4). Find an expression for $f_W(w)$, where $W = |X|$.

$$\begin{aligned} P(W \leq w) &= P(|X| \leq w) && \text{(plugging in for } W\text{)} \\ &= \begin{cases} P(-w \leq X \leq w) & \text{if } w \in [0, 2) \\ P(X \leq w) & \text{if } w \in [2, 4) \end{cases} && \text{(isolating } X\text{)} \\ &= \begin{cases} F_X(w) - F_X(-w) & \text{if } w \in [0, 2) \\ F_X(w) & \text{if } w \in [2, 4) \end{cases} && \text{(by our properties of CDFs)} \\ f_W(w) &= \begin{cases} f_X(w) + f_X(-w) & \text{if } w \in [0, 2) \\ f_X(w) & \text{if } w \in [2, 4) \end{cases} && \text{(differentiating w.r.t. } u\text{)} \end{aligned}$$

In this case, we have to pay close attention to which values in X map to which values in W. For values in $(-2, 2]$, the support of X is "folded," with two values of X mapping to a value in W. For values in $(2, 4)$, however, only one value in X is mapped to W.

(g) Theorem. Suppose we have a continuous random variable Y, and a $U = g(Y)$ is a strictly increasing or strictly decreasing function of Y. Then the PDF of U is given by

$$f_U(u) = f_Y\big(g^{-1}(u)\big) \left| \frac{dg^{-1}(u)}{du} \right|$$

(h) Aside. This follows directly from the method outlined above:

- If $g(Y)$ is increasing, then $P(g(Y) \leq u) = P(Y \leq g^{-1}(u)) = F_Y\big(g^{-1}(u)\big)$ and $\frac{dg^{-1}(u)}{du} > 0$
- If $g(Y)$ is decreasing, then $P(g(Y) \leq u) = P(Y \geq g^{-1}(u)) = 1 - F_Y\big(g^{-1}(u)\big)$ and $\frac{dg^{-1}(u)}{du} < 0$

This method can save a bit of time, however, if you're given a strictly increasing or decreasing function.

(i) Example. Suppose we have a random variable Y which measures tons of sugar refined per day. The distribution of Y is given by

$$f_Y(y) = 2y \qquad y \in [0,1]$$

Suppose it costs the company \$300 per ton to refine sugar, with fixed costs of \$100 per day. Then the daily profit in hundreds of dollars is $U = 3Y - 1$. Find the PDF of U

$$\begin{aligned} U = g(Y) &= 3Y - 1 && \text{(the transformation)} \\ Y = g^{-1}(U) &= \frac{U+1}{3} && \text{(solving for } Y\text{)} \\ \frac{\partial g^{-1}(U)}{\partial U} &= \frac{1}{3} && \text{(differentiating w.r.t. } U\text{)} \\ f_U(u) &= 2\left(\frac{u+1}{3}\right) \cdot \left|\frac{1}{3}\right| && \text{(employing our theorem)} \\ &= \frac{2}{9}(u+1) \qquad u \in (-1,2) \end{aligned}$$

(j) Aside. We will probably be required to know several transformations by heart, which many students have seen before. While we're getting ahead of ourselves a bit (some of these are multivariate transformations), these are worth knowing:

i. If $X \sim N(\mu, \sigma^2)$, then $\left(\frac{X-\mu}{\sigma}\right) \sim N(0,1)$

ii. If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_{(1)}$

iii. If $Y \sim \chi^2_{(k)}$ and $W \sim \chi^2_{(v)}$ are independent, then $Y + W \sim \chi^2_{(k+v)}$

iv. If $Z \sim N(0,1)$ and $Y \sim \chi^2_{(k)}$ are independent, then $\frac{Z}{\sqrt{Y/k}} \sim t_{(k)}$

v. If $U_1 \sim \chi^2_{(k_1)}$ and $U_2 \sim \chi^2_{(k_2)}$ are independent, then $\frac{U_1/k_1}{U_2/k_2} \sim F_{(k_1,k_2)}$

Note once again the assumption of independence, which we will dicuss in depth shortly.

8. Multiple Random Variables

(a) Definition. Let (X, Y) be a discrete, bivariate, random vector. Then the function $f_{XY}(x,y) : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f_{XY}(x,y) = P(X = x, Y = y)$$

is the **joint probability mass function**.

(b) Example. Consider the following table with associated probabilities for discrete random variables X and Y, where each may take on values in the set $\{1,2,3\}$:

		X		
		1	2	3
	1	0	1/8	1/4
Y	2	1/12	1/4	0
	3	1/6	1/8	0

This is a table representation of a joint PMF, which could be written out as a nine-element piecewise function. Each cell contains the probability $P(X = x_i, Y = y_j)$, e.g., $P(X = 1, Y = 1) = 0$ and $P(X = 3, Y = 1) = 1/4$.

(c) Definition. Given a discrete bivariate PMF $f_{XY}(x, y)$, the **marginal PMFs** of X and Y, denoted $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{XY}(x, y) \qquad \text{and} \qquad f_Y(y) = \sum_{x \in \mathbb{R}} f_{XY}(x, y)$$

(d) Example. Consider the distribution from the preceding example. To find the marginal PMF of Y, we sum across the rows:

		X			
		1	2	3	$f_Y(y)$
Y	1	0	1/8	1/4	3/8
	2	1/12	1/4	0	1/3
	3	1/6	1/8	0	7/24

$$f_Y(y) = \begin{cases} 3/8 \text{ if } Y = 1 \\ 1/3 \text{ if } Y = 2 \\ 7/24 \text{ if } Y = 3 \end{cases}$$

Analagously, to find the marginal PMF of X, we would sum over the values in each column.

(e) Definition. If (X, Y) is a continuous, bivariate, random vector, then $f_{XY}(x, y)$ is the **joint probability density function** if for every $A \in \mathbb{R}^2$:

$$P\big\{(X, Y) \in A\big\} = \int_A \int f_{XY}(x, y)\, dx\, dy$$

(f) Example. The bivariate uniform PDF, where $U_1 \in [0, 1]$ and $U_2 \in [0, 1]$, is given by

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0, 1] \text{ and } y \in [0, 1] \\ 0 & \text{else} \end{cases}$$

(g) Definition. Given a continuous bivariate PDF $f_{XY}(x, y)$, the **marginal PDFs** of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy \qquad \text{and} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx$$

(h) Example. Consider the joint PDF

$$f_{XY}(x, y) = \begin{cases} e^{-y} & \text{if } 0 < x < y < \infty \\ 0 & \text{else} \end{cases}$$

Then the marginal PDF of X can be found:

$$\begin{aligned} f_X(x) &= \int_x^{\infty} e^{-y} dy && \text{(integrating out } Y\text{)} \\ &= -e^{-y}\Big|_x^{\infty} && \text{(taking the integral)} \\ &= 0 - \big(-e^{-x}\big) && \text{(evaluating)} \\ f_X(x) &= e^{-x}, \quad x \in (0, \infty) \end{aligned}$$

(i) Definition. Let (X, Y) be a continuous (discrete) bivariate random vector with joint PDF (PMF) $f_{XY}(x, y)$ and margain PDFs (PMFs) $f_X(x)$ and $f_Y(y)$. Then for any x such that $f_X(x) > 0$, the **conditional PDF (PMF)** of Y given $X = x$ is given by

$$f(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

(j) Example. Given the joint PDF $f_{XY}(x, y) = e^{-y}$, where $0 < x < y < \infty$, find the conditional distribution of Y given $X = x$.

$$f_X(x) = e^{-x} \qquad \text{(from the previous example)}$$

$$f(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} \qquad \text{(by definition)}$$

$$= \frac{e^{-y}}{e^{-x}} \qquad \text{(plugging in the PDFs)}$$

$$f(y|x) = e^{-(y-x)} \quad y \geq x \qquad \text{(simplifying)}$$

We need to make sure that we specify the support $y \geq x$, as the joint PDF only takes on positive values when $0 < x < y < \infty$.

(k) Definition. Let (X, Y) be a bivariate random vector with joint PDF or PMF $f_{XY}(x, y)$ and marginal PDFs or PMFs $f_X(x)$ and $f_Y(y)$. Then X and Y are **independent random variables** if for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

(l) Aside. This is the formal definition of independence. If we need to show two variables are *not* independent, we must appeal to his definition. To show independence, however, we can rely on a simpler theorem.

(m) Theorem. X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that for all $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f_{XY}(x, y) = g(x)h(y)$$

(n) Aside. This gives us a weaker condition to check, as it does not require the use of integrals or sums–we don't need to actually find the marginal distributions.

(o) Example. We can show that the RVs are independent for the joint PDF:

$$f_{XY}(x, y) = \frac{1}{384} x^2 y^4 e^{-y-(x/2)} \qquad x > 0, y > 0$$

Trying to integrate this might be difficult. Instead, we can use the theorem above:

$$\frac{1}{384} x^2 y^4 e^{-y-(x/2)} = \left(\frac{y^4 e^{-y}}{384}\right)\left(x^2 e^{-x/2}\right)$$

Because the PDF can be factored into two functions, one solely of X, and one solely of Y, X and Y are independent.

What about the PDF from before?

$$f_{XY}(x,y) = e^{-y} \quad 0 < x < y < \infty$$

Even though the PDF looks like it could be factored, we have dependence in the support. If we rewrite with an indicator function:

$$f_{XY}(x,y) = e^{-y}\mathbb{1}\{0 < x < y < \infty\}$$

We clearly can't factor the function. To rigorously prove that these variables are not independent however, we'd need to appeal to the definition.

(p) Aside. Note that we can extend the concepts above to more than two dimensions. For example, we can get a marginal distribution for a subset of n jointly distributed random variables by integrating/summing over the remaining (i.e., we could find the marginal PDF of $X_1, \ldots, X_k$ by integrating the joint PDF over $X_{k+1}, \ldots, X_n$). Similarly, we could find a conditional PDF, e.g. $f(y|x_1, x_2, \ldots, x_n)$, which may interest us later on in econometrics. For the purposes of the first year, however, you will rarely manipulate distributions with more than two random variables. There is at least one notable exception.

(q) Definition. Let $X_1, \ldots, X_n$ be random variables with joint PDF or PMf $f_{\mathbf{X}}(x_1, \ldots x_n)$ and let $f_{X_i}(x_i)$ denote the marginal PDF or PMF of X_i. Then if $X_1, \ldots, X_n$ are **mutually independent random variables** if for every $(x_1, \ldots, x_n)$

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

9. **Multivariate Moments**

(a) Definition. Expectations of functions of random vectors are analogous to the univariate case. For a real-valued function $g(x,y)$ defined on the support of a bivariate random vector (X,Y), the expectation of $g(X,Y)$ is

$$\mathbb{E}\big[g(X,Y)\big] = \sum_{x\in\mathbb{R}} \sum_{y\in\mathbb{R}} g(x,y) f_{XY}(x,y) \qquad \text{(if } (X,Y) \text{ is a discrete)}$$

$$\mathbb{E}\big[g(X,Y)\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\, dx\, dy \qquad \text{(if } (X,Y) \text{ is cotinuous)}$$

(b) Definition. Let Y conditional on $X = x$ follow the distribution $f(y|x)$. If $g(Y)$ is a real-valued function of Y, then the **conditional expectation** of $g(Y)$ given that $X = x$ is given by

$$\mathbb{E}\big[g(Y)|X = x\big] = \sum_{y\in\mathbb{R}} g(y) f(y|x) dy \qquad \text{(if } Y \text{ is discrete)}$$

$$\mathbb{E}\big[g(Y)|X = x\big] = \int_{-\infty}^{\infty} g(y) f(y|x) dy \qquad \text{(if } Y \text{ is continuous)}$$

(c) Aside. Note that $\mathbb{E}[Y|X = x]$ is a value, i.e., the mean of Y given that we observe $X = x$. We also frequently use the expected value $\mathbb{E}[Y|X]$, which is a random variable—we don't know what the mean is until we know the value of X. This is one of the relatively few times where making a sharp distinction between x and X is important to us.

(d) Theorem. If X and Y are any two random variables, then

$$\mathbb{E}\big[X\big] = \mathbb{E}\Big[\mathbb{E}\big[X|Y\big]\Big]$$

This is known as the **Law of Iterated Expectations**.

(e) Definition. Given the conditions above, the **conditional variance** of Y given $X = x$

$$\mathrm{Var}\big[Y|X=x\big] = \mathbb{E}\big[Y^2|X=x\big] - \mathbb{E}\big[Y|X=x\big]^2$$

(f) Definition.The **covariance** of X and Y is the number defined by

$$\mathrm{Cov}(X,Y) = \mathbb{E}\big[(X-\mu_x)(Y-\mu_y)\big]$$

(g) Aside. Note that we frequently employ a simpler formula, analagous to our alternative formula for the univariate variance:

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

(h) Definition.The **correlation** of X and Y is the number defined by

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

(i) Aside. Note that the correlation is always between -1 and 1, and is the "unit-less" version of the covariance, where $\rho = -1$ and $\rho = 1$ represent perfect linear relationships between X and Y. As always, note that correlations only measure *linear* relationships.

(j) Theorem. If X and Y are any two random variables and a and b are any two constants, then

$$\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X,Y)$$

(k) Theorem. If X and Y are independent random variables, then the following are satisfied:

i. If $g(x)$ is a function only of x and $h(y)$ is a function only of y, then

$$\mathbb{E}\big[g(X)h(Y)\big] = \mathbb{E}\big[g(X)\big]\mathbb{E}\big[h(Y)\big]$$

ii. $\mathrm{Cov}(X,Y) = 0$

(l) Aside. Note that independence implies these hold, but not the other way around. Pay particular attention to the fact that $\mathrm{Cov}(X,Y) = 0$ DOES NOT imply independence.

10. **Multivariate Transformations**

(a) Definition. Let X_1 and X_2 be continuous random variables with the joint PDF $f_{X_1X_2}(X_1, X_2)$. Further, suppose that for all (x_1, x_2) and $f_{x_1x_2}(x_1, x_2) > 0$, we have one-to-one transformations $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$. The following constitutes the **bivariate transformation method**.

- Find $x_1 = h_1(y_1, y_2)$ and $x_2 = h_2(y_1, y_2)$, where

$$h_1(y_1,y_2) = g_1^{-1}(y_1,y_2) \qquad \text{and} \qquad h_2(y_1,y_2) = g_2^{-1}(y_1,y_2)$$

- Find the partial derivatives of h_1 and h_2 w.r.t. y_1 and y_2 and define the Jacobian matrix

$$\mathbf{J} = \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} \end{bmatrix}$$

- Then the joint PDF of Y_1 and Y_2 is

$$f_{Y_1Y_2}(y_1, y_2) = f_{X_1X_2}\Big(h_1\big(y_1, y_2\big),\, h_2\big(y_1, y_2\big)\Big)\big|\det(\mathbf{J})\big|$$

(b) Aside. Note that almost all bivariate transformations you'll do in the first year will involve one-to-one transformations. The other type of multivariate transformation that you will probably do during the first year will involve finding the distribution of a linear combination of random variables. In particular, we're interested in the distribution of things like the mean or conditional mean. In those cases, using our theorems regarding MGFs (if the MGFs exists, which they do for nearly all distributions we use during the first year) is a MUCH easier way to complete problems.