QUICK REVIEW OF PSTAT 120A

PSTAT 120B: Mathematical Statistics, I
Summer Session A, 2024 with Instructor: Ethan P. Marzban

DISCLAIMER: This is not meant to be a comprehensive treatment of PSTAT 120A material! Rather, I hope you can use this as a starting point, and can subsequently refer to your own notes from PSTAT 120A.

Contents

1 Fundamentals of Probability

Recall that an experiment, simply put, is any procedure we can repeat an infinite number of times, where each time we repeat the procedure the same fixed set of "things" (called outcomes) can occur. The set of all outcomes of an experiment is called the outcome space, and is denoted Ω.

We can define an event to be a subet of Ω; in other words, events are just sets comprised of outcomes. The event space, denoted $\mathcal{F}$, is the set of all outcomes associated with an experiment. As such, mathematically, $\mathcal{F}$ is simply a collection of subsets of Ω.

Example 1

Consider the experiment of tossing a coin twice and recording the outcomes.

Letting (X, Y) denote "the first toss landed X and the second toss landed Y" and H denote "heads" and T denote tails, we can express the outcome space of this experiment as

$$\Omega = \{(H,H)\,,\,(H,T)\,,\,(T,H)\,,\,(T,T)\} =: \{H,T\}^2$$

The event $E_1 :=$ "we observe at least one head" can be expressed mathematically as

$$E_1 = \{(H,H)\,,\,(H,T)\,,\,(T,H)\} \subseteq \Omega$$

We may take the event space of this experiment to be the power set of Ω: $\mathcal{F} = 2^\Omega$

Now, notice how we use the word "probability" in our everyday speech: "the probability of winning the lottery is one in a million", or "the chance of rain tomorrow is 25%", etc. We can see that, in a sense, "probability" is a *function* that acts on *events*. Indeed, we can make this more formal:

Definition (Probability Measure)

Given an outcome space Ω and an event space $\mathcal{F}$, a function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is said to be a probability measure if it satisfies the following three properties (which are collectively often referred to as the Axioms of Probability):

(1) Nonnegativity: $(\forall E \in \mathcal{F})[\mathbb{P}(E) \geq 0]$

(2) Probability of the Outcome Space is Unity: $\mathbb{P}(\Omega) = 1$

(3) Countable Additivity: for $\{E_i\}_{i=1}^{\infty} \subseteq \Omega$ with $E_i \cap E_j = \varnothing$ for every $i \neq j$ (collections of events satisfying this condition are said to be parwise disjoint),

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

The quantity $(\Omega, \mathcal{F}, \mathbb{P})$ - i.e. a collection of an outcome space, event space, and probability measure - is called a probability space.

From the Axioms of Probability (AP), we can derive several useful and important results. As an exercise, I encourage you to prove these on your own.

Theorem (Probability Rules)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we have the following:

- **Complement Rule:** $(\forall E \in \mathcal{F})[\mathbb{P}(E^{\complement}) = 1 - \mathbb{P}(E)]$
- **Set Difference Rule:** $(\forall E, F \in \mathcal{F})[\mathbb{P}(E \setminus F) = \mathbb{P}(E) - \mathbb{P}(E \cap F)$
- **Addition Rule:** $(\forall E, F \in \mathcal{F})[\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$
- **Inclusion-Exclusion Rule:** for a collection of events $\{E_i\}_{i=1}^{n} \subseteq \Omega$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbb{P}(E_i \cap E_j \cap E_k) + \cdots + (-1)^{n-1} \sum_{i<\cdots<n} \mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right)$$

Example 2

A cinema recently conducted a survey and found that 80% of moviegoers purchase popcorn, 60% purchase a drink, and 50% purchase both popcorn and a drink (or both).

Defining $P :=$ "randomly selected moviegoer purchases popcorn" and $D :=$ "randomly selected

moviegoer purchases a drink", the problem tells us that

$$\mathbb{P}(P) = 0.8; \quad \mathbb{P}(D) = 0.6; \quad \mathbb{P}(P \cap D) = 0.5$$

With this, we can compute the probability that a randomly-selected moviegoer purchases either popcorn or a drink (or both) to be

$$\mathbb{P}(P \cup D) = \mathbb{P}(P) + \mathbb{P}(D) - \mathbb{P}(P \cap D) = 0.8 + 0.6 - 0.5 = 0.9 = 90\%$$

Similarly, the probability that a randomly-selected moviegoer purchases only popcorn (and not a drink) is

$$\mathbb{P}(P \setminus D) = \mathbb{P}(P) - \mathbb{P}(P \cap D) = 0.8 = 0.5 = 0.3 = 30\%$$

2 Conditional Probability and Independence

Often times, information about one event will influence our beliefs about another. For example, in the absence of any information, we might say "the chance of rain tomorrow is 50%." However, the knowledge that we are in Santa Barbara in the summer would likely *decrease* what we believe the true probability of rain tomorrow is.

This is precisely where conditional probabilities are useful - they give us a way to express an "updating of beliefs."

Definition (Conditional Probability Measure)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and events $E, F \in \mathcal{F}$ with $\mathbb{P}(F) > 0$, we define the **conditional probability of E given F** to be

$$\mathbb{P}_F(E) := \mathbb{P}(E \mid F) := \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}$$

A couple of notes:

- The notation $\mathbb{P}(E \mid F)$ is more common, though $\mathbb{P}_F(E)$ is still used fairly often.
- Note, crucially, that $\mathbb{P}(E \mid F)$ exists only when $\mathbb{P}(F) > 0$. If $\mathbb{P}(F) = 0$, then $\mathbb{P}(E \mid F)$ is simply undefined.
- $\mathbb{P}_F(E)$ is actually a valid probability measure; I encourage you to prove this as an exercise (all you need to do is show that all three of the Axioms of Probability are satisfied).
- One interpretation of the quantity $\mathbb{P}(E \mid F)$ is as our beliefs about the event E, updated to reflect the fact that F has occurred. In this way, conditional probabilities are "if-then" statements: $\mathbb{P}(E \mid F)$ answers the question "if F, what is the probability of E?"

Definition (Partition)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, an event $F \in \mathcal{F}$, and a collection of events $\{E_i\}_{i=1}^n$, we say that the collection $\{E_i\}_{i=1}^n$ partitions (or forms a partition of) F if:

(1) the sequence is pairwise disjoint (i.e. $E_i \cap E_j = \varnothing$ for every $i \neq j$)

(2) $\bigcup_{i=1}^{n} E_i = F$

Theorem (Law of Total Probability)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, an event $F \in \mathcal{F}$, and a partition $\{E_i\}_{i=1}^{\infty}$ of Ω,

$$\mathbb{P}(F) = \sum_{i=1}^{\infty} \mathbb{P}(F \mid E_i) \cdot \mathbb{P}(E_i)$$

The Law of Total Probability (LoTP) is particularly useful in that it allows us to decompose an unconditional probability into a sum of terms involving conditional probabilities.

Theorem (Bayes' Rule)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and events $E, F \in \mathcal{F}$ with $\mathbb{P}(E) \neq 0$ and $\mathbb{P}(F) \neq 0$,

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(F \mid E) \cdot \mathbb{P}(E)}{\mathbb{P}(F)}$$

Bayes' rule effectively allows us to "reverse the order of a conditional probability." It is common to use the LoTP to compute the denominator of Bayes' Rule.

Example 3

Suppose 5% of a particular population has been infected with a disease. There exists a test for this disease, but it is not perfect: in 10% of cases the test incorrectly reports a healthy person as having the disease, and in 7% of cases the test incorrectly reports a diseased person as being healthy.

Let $D :=$ {a randomly-selected person has the disease} and $T_+ :=$ {a randomly-selected person tests positive}. Then, from the problem statement,

$$\mathbb{P}(D) = 0.05; \qquad \mathbb{P}(T_+ \mid D^{\complement}) = 0.1; \qquad \mathbb{P}(T_+^{\complement} \mid D) = 0.07$$

The complement rule allows us to compute

$$\mathbb{P}(D^{\complement}) = 0.95; \qquad \mathbb{P}(T_+^{\complement} \mid D^{\complement}) = 0.9; \qquad \mathbb{P}(T_+ \mid D) = 0.93$$

By the Law of Total Probability, the probability that a randomly-selected person tests positive is

$$\begin{aligned} \mathbb{P}(T_+) &= \mathbb{P}(T_+ \mid D) \cdot \mathbb{P}(D) + \mathbb{P}(T_+ \mid D^{\complement}) \cdot \mathbb{P}(D^{\complement}) \\ &= (0.93) \cdot (0.05) + (0.1) \cdot (0.95) = 0.1415 \end{aligned}$$

Thus, by Bayes' Rule, the probability that a randomly-selected person has the disease given that they have tested positive is

$$\mathbb{P}(D \mid T_+) = \frac{\mathbb{P}(T_+ \mid D) \cdot \mathbb{P}(D)}{\mathbb{P}(T_+)} = \frac{(0.93) \cdot (0.05)}{0.1415} \approx 0.3286 = 32.86\%$$

Definition (Independence)

Given a probability space $(\Omega, \mathbb{P}, \mathcal{F})$, events $E, F \in \mathcal{F}$ are said to be independent, notated $E \perp F$, if any of the following (equivalent) statements hold:

1. $\mathbb{P}(E \mid F) = \mathbb{P}(E)$
2. $\mathbb{P}(F \mid E) = \mathbb{P}(F)$
3. $\mathbb{P}(E \cap F) = \mathbb{P}(E) \cdot \mathbb{P}(F)$

Here's one way to interpret this definition: take a look at condition (1): $\mathbb{P}(E \mid F)$. In a sense, $\mathbb{P}(E \mid F)$ represents our beliefs about the event E in the presence of knowledge that F has occurred. The statement $\mathbb{P}(E \mid F) = \mathbb{P}(E)$ is then asserting that our beliefs about the event E in the presence of knowledge that F has occurred *have remain unchanged* from our beliefs about E without any information on F. In this way, we are asserting that the events E and F "don't affect each other-" i.e. that they are "independent".

Example 2 (cont'd)

Consider the following situation again: A cinema recently conducted a survey and found that 80% of moviegoers purchase popcorn, 60% purchase a drink, and 50% purchase both popcorn and a drink (or both).

Note that

$$\mathbb{P}(P) \cdot \mathbb{P}(D) = (0.8) \cdot (0.6) = 0.48 \neq 0.5 = \mathbb{P}(P \cap D)$$

Hence, the events P and D are not independent: customers do *not* appear to purchase popcorn and drinks independently.

Independence of more than two events is a bit more complicated:

Definition (Mutual Independence)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a sequence $\{E_i\}_{i=1}^n$ are said to be **mutually independent** if for every subsequence $\{E_{i_1}, \cdots, E_{i_k}\} \subseteq \{E_i\}_{i=1}^n$, with $2 \leq k \leq n$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$, it holds that

$$\mathbb{P}\left(\bigcap_{j=1}^{k} E_{i_j}\right) = \prod_{i=1}^{k} \mathbb{P}(E_{i_j})$$

For example, to establish the independence of 4 events A, B, C, D, we would need to verify all of the following:

- $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$
- $\mathbb{P}(A \cap C) = \mathbb{P}(A) \cdot \mathbb{P}(C)$
- $\mathbb{P}(A \cap D) = \mathbb{P}(A) \cdot \mathbb{P}(D)$
- $\mathbb{P}(B \cap C) = \mathbb{P}(B) \cdot \mathbb{P}(C)$
- $\mathbb{P}(B \cap D) = \mathbb{P}(B) \cdot \mathbb{P}(D)$

- $\mathbb{P}(C \cap D) = \mathbb{P}(C) \cdot \mathbb{P}(D)$
- $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C)$
- $\mathbb{P}(A \cap B \cap D) = \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(D)$
- $\mathbb{P}(A \cap C \cap D) = \mathbb{P}(A) \cdot \mathbb{P}(C) \cdot \mathbb{P}(D)$
- $\mathbb{P}(B \cap C \cap D) = \mathbb{P}(B) \cdot \mathbb{P}(C) \cdot \mathbb{P}(D)$
- $\mathbb{P}(A \cap B \cap C \cap D) = \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C) \cdot \mathbb{P}(D)$

One can show that, to prove the independence of n events, we must check $2^n - n - 1$ conditions.

3 Random Variables

We define a random variable X to be a function that maps from Ω to $\mathbb{R}$. The support or state space of a random variable X, notated S_X, is defined to be $X(\Omega)$.

Example 4

Consider the experiment of tossing two fair coins, and recording the number of heads observed. Let X denote the number of heads observed.

We have previously seen that one way to express the outcome space of this experiment is

$$\Omega = \{H, T\}^2 = \{(H, H)\,,\ (H, T)\,,\ (T, H)\,,\ (T, T)\}$$

Since X counts the number of heads, we see that

$$X((H, H)) = 2; \qquad X((H, T)) = 1; \qquad X((T, H)) = 1; \qquad X((T, T)) = 0$$

Hence, $S_X = \{0, 1, 2\}$.

Random variables are classified by their support. If S_X is at most countable (i.e. finite or countably infinite), then X is said to be a discrete random variable. If S_X is uncountable, X is said to be a continuous random variable. It is possible for a random variable to be neither discrete nor continuous; such distributions are often called mixed random variables.

Example 4 (cont'd)

Since $S_X = \{0, 1, 2\}$ is a finite set, the random variable X is discrete.

Discrete random variables are characterized by a probability mass function (PMF, or just "mass function" for short), defined as $p_X(x) := \mathbb{P}(X = x)$. Indeed, a valid PMF must crucially satisfy two properties:

(1) Nonnegativity: $(\forall x \in \mathbb{R})[0 \leq p_X(x) \leq 1]$

(2) Summing to Unity: $\sum_{\text{all } x} p_X(x) = 1$

Example 4 (cont'd)

$$p_X(2) := \mathbb{P}(X = 2) = \mathbb{P}(\{(H, H)\}) = \frac{1}{4}$$
$$p_X(1) := \mathbb{P}(X = 2) = \mathbb{P}(\{(H, T) \cup (T, H)\}) = \frac{1}{2}$$
$$p_X(0) := \mathbb{P}(X = 2) = \mathbb{P}(\{(T, T)\}) = \frac{1}{4}$$

Therefore, we can summarize the PMF of X as

$$p_X(x) = \begin{cases} 1/4 & \text{if } x = 0 \\ 1/2 & \text{if } x = 1 \\ 1/4 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

or tabularly as

$\boldsymbol{x}$	$\mathbf{0}$	$\mathbf{1}$	$\mathbf{2}$
$\mathbb{P}(\boldsymbol{X = x})$	$1/4$	$1/2$	$1/4$

Note

The notation $\{X = x\}$ is really a shorthand:

$$\{X = x\} := \{\omega \in \Omega : X(\omega) = x\}$$

Similarly,

$$\{X \in B\} := \{\omega \in \Omega : X(\omega) \in B\}$$

For example,

$$\{X \leq x\} := \{\omega \in \Omega : X(\omega) \leq x\}$$

Given a PMF, several quantities can be computed:

(1) Expectation: $\mathbb{E}[X] := \sum_x x \cdot p_X(x)$

(2) Law of the Unconscious Statistician (LOTUS): $\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$

(3) n^{th} moment: $\mu_n^{(X)} := \mathbb{E}[X^n]$

(4) Variance: $\text{Var}(X) := \mathbb{E}\left\{(X - \mathbb{E}[X])^2\right\} = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

(5) Cumulative Mass Function (CMF): $F(x) := \mathbb{P}(X \leq x)$

Continuous random variables are characterized by a probability density function (PDF, or just "density function" for short), which is a function $f_X : \mathbb{R} \to \mathbb{R}$ satisfying

(1) **Nonnegativity:** $(\forall x \in \mathbb{R})[0 \leq f_X(x) \leq 1]$

(2) **Integrating to Unity:** $\int_{\mathbb{R}} f_X(x)\, \mathrm{d}x = 1$

Given a PDF, several quantities can be computed:

(1) Expectation: $\mathbb{E}[X] := \int_{\mathbb{R}} x f_X(x)\, \mathrm{d}x$

(2) Law of the Unconscious Statistician (LOTUS): $\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) \cdot f_X(x)\, \mathrm{d}x$

(3) n^{th} moment: $\mu_n^{(X)} := \mathbb{E}[X^n]$

(4) Variance: $\mathrm{Var}(X) := \mathbb{E}\left\{(X - \mathbb{E}[X])^2\right\} = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

(5) Cumulative Distribution Function (CMF): $F(x) := \mathbb{P}(X \leq x)$

Additionally, we define the survival function of a random variable X as $\overline{F_X}(x) := \mathbb{P}(X > x) = 1 - F_X(x)$.

Definition (Indicator Random Variable)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an event $E \in \mathcal{F}$, we define the indicator random variable of the event E (often referred to simply as the "indicator") to be

$$\mathbb{1}_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \in E^{\complement} \end{cases}$$

Notationally, it is common to drop the ω and write

$$\mathbb{1}_E := \begin{cases} 1 & \text{if } E \\ 0 & \text{if } E^{\complement} \end{cases}$$

Indicators have a few important and useful properties:

- $\mathbb{E}[\mathbb{1}_E] = \mathbb{P}(E)$
- $\mathrm{Var}(\mathbb{1}_E) = \mathbb{P}(E) \cdot \mathbb{P}(E^{\complement})$
- $\mathbb{1}_E \cdot \mathbb{1}_F = \mathbb{1}_{E \cap F}$

Example 5

Consider a random variable X with density given by

$$f_X(x) = \frac{1}{\theta} \cdot e^{-x/\theta} \cdot 1_{\{x \geq 0\}}$$

where $\theta > 0$ is a deterministic (i.e. non-random) constant.

First note that the indicator tells us that $f_X(x) > 0$ only when $x \geq 0$; in other words, the support of X is $S_X = [0, \infty)$.

This is a valid density function because:

(1) **Nonnegativity:** Whenever $x < 0$ we have $f_X(x) = 0$. Whenever $x \geq 0$ we know $e^{-x/\theta} > 0$, and since $\theta > 0$ we have $f_X(x) > 0$. Thus, for any $x \in \mathbb{R}$, we have $f_X(x) \geq 0$.

(2) Integrating to Unity:

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{\theta}\cdot e^{-x/\theta}\cdot 1_{\{x\geq 0\}}\,\mathrm{d}x$$
$$= \int_0^{\infty} \frac{1}{\theta}e^{-x/\theta}\,\mathrm{d}x = \frac{1}{\theta}\cdot\theta\cdot e^{-x/\theta}\Big]_{x=\infty}^{x=0} = 1\ \checkmark$$

4 Distributions

Certain mass/density functions arise so often that it becomes useful to give them a name. We can essentially think of a distribution as a "package," giving us information about a PMF/PDF, CMF/CDF, support, expectation, etc. (We'll actually discuss the notion of distributions a bit further during the first few lectures of PSTAT 120B.)

For example, the density in Example 4 above is the density of the Exponential Distribution, which admits a single parameter θ. In general, we can consider parameters to be deterministic (i.e. nonrandom) constants that affect the shape/behavior of the mass/density function.

Often times, I find it useful to consider a "story" behind each distribution. That is, I encourage you to think of some typical examples that lead naturally to the different distributions - this will help you identify what distribution a particular random variable follows.

4.1 Important Discrete Distributions

Bernoulli Distribution: $X \sim \mathsf{Bern}(p)$, where $0 < p < 1$

- **Support**: $S_X = \{0, 1\}$
- **PMF**: $p_X(x) = p^x(1-p)^{1-x}$
- **Expectation/Variance**: $\mathbb{E}[X] = p; \quad \mathsf{Var}(X) = p(1-p)$
- **Story**: result of a single coin flip.

Binomial Distribution: $X \sim \mathsf{Bin}(n, p)$, where $n \in \mathbb{N}$ and $0 < p < 1$.

- **Support**: $S_X = \{0, 1, \cdots, n\}$
- **PMF**: $p_X(x) = \binom{n}{x}p^x(1-p)^{n-x}$
- **Expectation/Variance**: $\mathbb{E}[X] = np; \quad \mathsf{Var}(X) = np(1-p)$
- **Story**: number of heads in n tosses of a $p-$coin [i.e. a coin that lands heads with probability p]

Geometric Distribution on $\{0, 1, \cdots\}$: $X \sim \text{Geom}(p)$ on $\{0, 1, \cdots\}$, where $0 < p < 1$.

- **Support**: $S_X = \{0, 1, \cdots\}$
- **PMF**: $p_X(x) = (1-p)^x p$
- **Expectation/Variance**: $\mathbb{E}[X] = \frac{1-p}{p}; \quad \text{Var}(X) = \frac{1-p}{p^2}$
- **Story**: number of failures before the first success, where there is a p probability of success on any given trial.

Geometric Distribution on $\{1, 2, \cdots\}$: $X \sim \text{Geom}(p)$ on $\{1, 2, \cdots\}$, where $0 < p < 1$.

- **Support**: $S_X = \{1, 2, \cdots\}$
- **PMF**: $p_X(x) = (1-p)^{x-1} p$
- **Expectation/Variance**: $\mathbb{E}[X] = \frac{1}{p}; \quad \text{Var}(X) = \frac{1-p}{p^2}$
- **Story**: number of total trials before the first success, where there is a p probability of success on any given trial.

Negative Binomial Distribution on $\{0, 1, \cdots\}$: $X \sim \text{NegBin}(r, p)$ on $\{0, 1, \cdots\}$, where $r \in \mathbb{N}$ and $0 < p < 1$.

- **Support**: $S_X = \{0, 1, \cdots\}$
- **PMF**: $p_X(x) = \binom{x+r-1}{x}(1-p)^x p^r$
- **Expectation/Variance**: $\mathbb{E}[X] = \frac{r(1-p)}{p}; \quad \text{Var}(X) = \frac{r(1-p)}{p^2}$
- **Story**: number of failures before the r^{th} success, where there is a p probability of success on any given trial.

Negative Binomial Distribution on $\{r, r+1, \cdots\}$: $X \sim \text{NegBin}(r, p)$ on $\{r, r+1, \cdots\}$, where $r \in \mathbb{N}$ and $0 < p < 1$.

- **Support**: $S_X = \{r, r+1, \cdots\}$
- **PMF**: $p_X(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$
- **Expectation/Variance**: $\mathbb{E}[X] = \frac{r}{p}; \quad \text{Var}(X) = \frac{r(1-p)}{p^2}$
- **Story**: number of total trials before the r^{th} success, where there is a p probability of success on any given trial.

Poisson Distribution: $X \sim \text{Pois}(\lambda)$, where $\lambda > 0$

- **Support**: $S_X = \{0, 1, \cdots\}$
- **PMF**: $p_X(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$
- **Expectation/Variance**: $\mathbb{E}[X] = \lambda; \quad \text{Var}(X) = \lambda$
- **Story**: number of arrivals in an interval of time.

Hypergeometric Distribution: $X \sim \text{HyperGeom}(N, G, n)$, were $N \in \{0, 1, \cdots,\}$, $G \in \{0, 1, \cdots, N\}$, and $n \in \{0, 1, \cdots, N\}$

- **Support**: $S_X = \{\text{max}\{0\ ,\ n + G - N\}, \cdots, \text{min}\{n\ ,\ G\}\}$
- **PMF**: $p_X(x) = \frac{\binom{G}{x}\binom{N-G}{n-x}}{\binom{N}{n}}$
- **Expectation/Variance**: $\mathbb{E}[X] = n \cdot \frac{G}{N}; \quad \text{Var}(X) = n \cdot \left(\frac{G}{N}\right) \cdot \left(1 - \frac{G}{N}\right) \cdot \frac{N-n}{N-1}$
- **Story**: number of "good" elements in a sample of size n, drawn from a lot of N total elements of which G are "good".

Discrete Uniform Distribution: $X \sim \text{DiscUnif}\{x_1, \cdots, x_n\}$

- **Support**: $S_X = \{x_1, \cdots, x_n\}$
- **PMF**: $p_X(x) = \frac{1}{n}$, where $n = |\{x_1, \cdots, x_n\}|$
- **Expectation/Variance**: $\mathbb{E}[X] = \overline{x}_n := \frac{1}{n}\sum_{i=1}^n x_i; \quad \text{Var}(X) = \frac{1}{n}\sum_{i=1}^n x_i^2 - (\overline{x}_n)^2$
- **Story**: outcome of drawing a number at random from the set $\{x_1, \cdots, x_n\}$.

4.2 Important Continuous Distributions

Uniform Distribution: $X \sim \text{Unif}[a, b]$, where $-\infty < a < b < \infty$.

- **Support**: $S_X = [a, b]$
- **PMF**: $f_X(x) = \frac{1}{b-a} \cdot \mathbb{1}_{\{x \in [a,b]\}}$
- **Expectation/Variance**: $\mathbb{E}[X] = \frac{a+b}{2}; \qquad \text{Var}(X) = \frac{(b-a)^2}{12}$

Exponential Distribution: $X \sim \text{Exp}(\theta)$, $\theta > 0$

- **Support**: $S_X = [0, \infty)$
- **PMF**: $f_X(x) = \frac{1}{\theta}e^{-x/\theta} \cdot \mathbb{1}_{\{x \geq 0\}}$
- **Expectation/Variance**: $\mathbb{E}[X] = \theta; \quad \text{Var}(X) = \theta^2$

Note: the parameterization of the exponential distribution listed above is a little different than the one typically used in PSTAT 120A. Please be aware that this new parameterization is the one we will use in PSTAT 120B.

Normal Distribution: $X \sim \mathcal{N}(\mu, \sigma^2)$

- **Support**: $S_X = \mathbb{R}$
- **PMF**: $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$
- **Expectation/Variance**: $\mathbb{E}[X] = \mu; \qquad \text{Var}(X) = \sigma^2$

Definition (Standard Normal CDF)

If $X \sim \mathcal{N}(0, 1)$ [which is called the standard normal distribution], its CDF is defined to be the function $\Phi(\cdot)$. In other words:

$$\Phi(x) := \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, \mathrm{d}z$$

Theorem (Standardization of Normal Distributions)

Given $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$Z := \left(\frac{X-\mu}{\sigma}\right) \sim \mathcal{N}(0, 1)$$

We'll actually revisit this result partway through PSTAT 120B.

4.3 Generating Functions

It turns out that there is another class of functions (aside from PMFs/PDFs and CMFs/CDFs) that can be used to describe distributions - these are called generating functions. In PSTAT 120A, you were (hopefully) introduced to the moment-generating function (MGF), defined as

$$M_X(t) := \mathbb{E}[e^{tX}]$$

Note that MGFs are computed as sums if X is discrete and integrals if X is continuous. (We won't talk about MGFs of mixed distributions in this class.)

MGFs are named the way they are because of the following useful property:

Theorem

Given a random variable X with MGF $M_X(t)$ that is finite in an interval around the origin, we have

$$M_X^{(n)}(0) = \mathbb{E}[X^n]$$

That is, the n^{th} moment of X can be computed by taking the n^{th} derivative of the MGF and evaluating the result at 0.

Distribution	$M_{\boldsymbol{X}}(t)$
Binomial	$M_X(t) = (1 - p + pe^t)^n$
Geometric on $\{0, 1, \cdots\}$	$M_X(t) = \begin{cases} \frac{p}{1-(1-p)e^t} & \text{if } t < -\ln(1-p) \\ \infty & \text{otherwise} \end{cases}$
Geometric on $\{1, 2, \cdots\}$	$M_X(t) = \begin{cases} \frac{pe^t}{1-(1-p)e^t} & \text{if } t < -\ln(1-p) \\ \infty & \text{otherwise} \end{cases}$
Negative Binomial on $\{0, 1, \cdots\}$	$M_X(t) = \begin{cases} \left(\frac{p}{1-(1-p)e^t}\right)^r & \text{if } t < -\ln(1-p) \\ \infty & \text{otherwise} \end{cases}$
Negative Binomial on $\{1, 2, \cdots\}$	$M_X(t) = \begin{cases} \left(\frac{pe^t}{1-(1-p)e^t}\right)^r & \text{if } t < -\ln(1-p) \\ \infty & \text{otherwise} \end{cases}$
Poisson	$M_X(t) = e^{\lambda(e^t-1)}$
Uniform	$M_X(t) = \begin{cases} \frac{e^{tb}-e^{ta}}{t(b-a)} & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$
Exponential	$M_X(t) = \begin{cases} (1-\theta t)^{-1} & \text{if } t < 1/\theta \\ 1 & \text{if } t = 0 \end{cases}$
Normal	$M_X(t) = \exp\left\{\mu t + \frac{\sigma^2}{2}t^2\right\}$

MGFs are not the only type of generating function. Another very popular generating function is the so-called probability generating function (PGF), defined as

$$G_X(z) := \mathbb{E}[z^X]$$

PGFs are not typically discussed in PSTAT 120A (or PSTAT 120B), but they arise very frequently in PSTAT 160A.

5 Random Vectors

It's often desirable to consider collections of random variables. This is typical of a situation in which outcomes are n-tuples of numbers. In such a situation, we define a random vector to be

$$\vec{\boldsymbol{X}} := \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} : \Omega \to \mathbb{R}^n$$

Continuous Random Vectors are described by a joint density function $f_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$, sometimes abbreviated $f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})$, that satisfies:

(1) Nonnegativity: $(\forall \vec{\boldsymbol{x}} \in \mathbb{R}^n)[0 \leq f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}}) \leq 1]$

(2) Integrating to Unity: $\int_{\mathbb{R}^n} f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})\, \mathrm{d}\vec{\boldsymbol{x}} = 1$

Note that $\int_{\mathbb{R}^n} f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})\,\mathrm{d}\vec{\boldsymbol{x}}$ is a shorthand notation for an n-dimensional integral:

$$\int_{\mathbb{R}^n} f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})\,\mathrm{d}\vec{\boldsymbol{x}} := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1,X_2,\cdots,X_n}(x_1,x_2,\cdots,x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\cdots\,\mathrm{d}x_n$$

The multivariate LOTUS (Law of the Unconscious Statistician) becomes

$$\mathbb{E}[g(\vec{\boldsymbol{X}})] = \int_{\mathbb{R}^n} g(\vec{\boldsymbol{x}})\cdot f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})\,\mathrm{d}\vec{\boldsymbol{x}}$$

For example, given a bivariate random vector (X,Y) with joint density $f_{X,Y}(x,y)$, we can compute

$$\mathbb{E}[\cos(X^2-2X\sqrt{Y})] = \iint_{\mathbb{R}^2} \cos(x^2-2x\sqrt{y}) f_{X,Y}(x,y)\,\mathrm{d}A$$

Note

In this class, we'll often use notation like $(X,Y)\sim f_{X,Y}$ to be a shorthand for "the bivariate random vector (X,Y) has joint density $f_{X,Y}(\cdot,\cdot)$."

Definition (Independent Random Variables)

Random variables $X_1,\cdots,X_n$ are said to be independent if their joint density factors as a product of their marginals:

$$f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}}) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

For example, given $(X,Y)\sim f_{X,Y}$, we can conclude $X\perp Y$ if and only if $f_{X,Y}(x,y) = f_X(x)\cdot f_Y(y)$. An interesting consequence of this is that if the joint support of (X,Y) is nonrectangular, then we can automatically conclude $X\perp Y$. The converse is not necessarily true - just because the joint support *is* rectangular, we *cannot* automatically conclude that $X\perp Y$.

Probabilities are computed as integrals of the joint:

$$\mathbb{P}(\vec{\boldsymbol{X}}\in A) = \int_A f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})\,\mathrm{d}\vec{\boldsymbol{x}}$$

Recall from PSTAT 120A that, when computing probabilities in the multivariate setting, sketching the region of integration is absolutely crucial.

Example 5

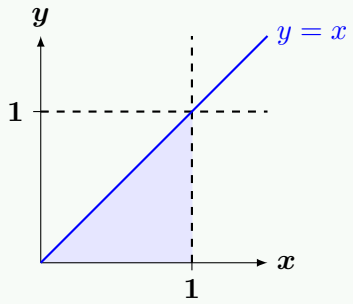
Let $(X,Y)\sim f_{X,Y}$ where

$$f_{X,Y}(x,y) = 8xy\cdot\mathbb{1}_{\{0\le y<x\le 1\}}$$

Here, the joint support is the set

$$S_{(X,Y)} := \{(x,y)\in\mathbb{R}^2 : 0\le y<x\le 1\}$$

which, when sketched, looks like:



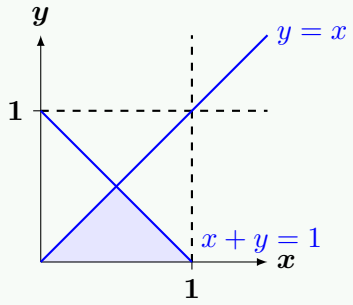
Suppose we want to compute $\mathbb{P}(X+Y<1)$. We can accomplish this by computing

$$\iint_{\mathcal{R}} f_{X,Y}(x,y)\,\mathrm{d}A$$

where $\mathcal{R}$ is the region

$$\mathcal{R} := \{(x,y) \in \mathbb{R}^2 : x+y<1\} \cap S_{X,Y}$$

which, when sketched, looks like



Notice that if we integrated in the order $\mathrm{d}y\,\mathrm{d}x$ we would have to split our integral into two. This is not the case if we use the order $\mathrm{d}x\,\mathrm{d}y$; as such, let's use this order of integration:

$$\begin{aligned}
\mathbb{P}(X+Y<1) &= \iint_{\mathcal{R}} f_{X,Y}(x,y)\,\mathrm{d}A \\
&= \int_0^{1/2}\int_y^{1-y} 8xy\,\mathrm{d}x\,\mathrm{d}y \\
&= 4\int_0^{1/2} y[(1-y)^2 - y^2]\,\mathrm{d}y \\
&= 4\int_0^{1/2} (y-2y^2)\,\mathrm{d}y = 4\left(\frac{1}{2}\cdot\frac{1}{4} - \frac{2}{3}\cdot\frac{1}{8}\right) = \frac{1}{6}
\end{aligned}$$

It's sometimes desired to quantify how related two random variables are. This is where the notions of covariance and correlation become useful.

Definition (Covariance and Correlation)

Given two random variables X and Y, we define their covariance to be

$$\text{Cov}(X,Y) := \mathbb{E}\left\{(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])\right\} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Because covariances are unbounded (i.e. range between $-\infty$ and ∞), we define the correlation to be a standardized version of covariance:

$$\text{Corr}(X,Y) := \frac{\text{Cov}(X,Y)}{\text{SD}(X) \cdot \text{SD}(Y)} \in [-1,1]$$

When $|\text{Corr}(X,Y)|$ is close to 1 (i.e. when the correlation is close to -1 or 1), there is indication that the two variables are strongly related. Note that if $X \perp Y$, then $\text{Cov}(X,Y) = 0$. The converse is not true: just because $\text{Cov}(X,Y) = 0$ doesn't mean $X \perp Y$.

Theorem (Bilinearity of Covariance)

Given collections $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ of random variables and sequences $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ of constants,

$$\text{Cov}\left(\sum_{i=1}^n a_i X_i\,,\ \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j)$$

This property, coupled with the fact that $\text{Cov}(X,X) = \text{Var}(X)$, allows us to derive the formula

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2\sum_{i<j} a_i a_j \text{Cov}(X_i, X_j)$$

As you can imagine, with random vectors it can become a bit complicated to express dependency structures. As such, it's common to use variance-covariance matrices:

$$\text{Var}(\vec{\boldsymbol{X}}) = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1,X_2) & \cdots & \text{Cov}(X_1,X_n) \\ \text{Cov}(X_1,X_2) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_1,X_n) & \text{Cov}(X_2,X_n) & \cdots & \text{Var}(X_n) \end{bmatrix} =: \boldsymbol{\Sigma}$$

Note that, since covariance is a symmetric operator [$\text{Cov}(X_i,X_j) = \text{Cov}(X_j,X_i)$], the matrix $\boldsymbol{\Sigma}$ will be a symmetric matrix.

6 Inequalities

Sometimes we won't have access to the full distribution of a random variable. In certain cases, we can still provide bounds on various probabilities involving said random variable by invoking Markov's Inequality and Chebyshev's Inequality.

Theorem (Markov's Inequality)

Given a nonnegative random variable X [i.e. the support of X contains only nonnegative values)] and a positive constant $c > 0$,

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[X]}{c}$$

Theorem (Chebyshev's Inequality)

Given a random variable X with finite mean μ and finite variance σ^2, and for a positive constant $c > 0$,

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

Recall that, in certain cases, these inequalities can provide correct yet useless bounds. For instance, consider a nonnegative random variable X with mean $\mathbb{E}[X] = 2$. By Markov's Inequality,

$$\mathbb{P}(X \geq 1) \leq \frac{\mathbb{E}[X]}{1} = 2$$

However, this upper bound of 2 doesn't provide any meaningful information as we know $\mathbb{P}(X \geq 1) \leq 1$.

Additionally, it's important to consider the assumptions of these inequalities. For example, a random variable X with density

$$f_X(x) = \frac{2}{x^3} \cdot \mathbb{1}_{\{x \geq 1\}}$$

has finite expectation but infinite variance. Hence, we cannot apply Chebyshev's Inequality to bound probabilities involving X (though, in practice, we wouldn't ever really use Chebyshev's Inequality since we have access to the density and can therefore compute probabilities exactly).

Index