Module 1 Recitation Notes

ESE/EEO 306
Summer Session II 2022
Dan Waxman

Abstract

We'll summarize the key concepts of the current module in a few pages, then thoroughly work through four or five problems. This material comes from the class slides or the class textbook [1] unless otherwise noted.

1 Key Concepts

1.1 Probability Baiscs

We began by talking about *random* events — exactly what that means is a matter of philosophy, but we'll just consider these things unpredictable. The set of outcomes in an experiment is the *sample space*, often denoted Ω, and an event $A \subset \Omega$ is a subset of the sample space.

For somewhat technical reasons that occur when we move to continuous spaces, we require the definition of a σ*-algebra*[1]. Denoting the complement of an event A as A^c, σ-algebras are defined as follows:

Definition 1. A σ-algebra $\mathcal{F}$ on a sample space Ω is a collection of events which is closed under union and complements, and contains the empty set as a member. In symbols:

- If $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$ (**Closed under unions**)
- If $A \in \mathcal{F}$, then $A^\mathsf{c} \in \mathcal{F}$ (**Closed under complements**)
- We have $\emptyset \in \mathcal{F}$ (**Contains the empty set**)

Given some sample space Ω and σ-algebra $\mathcal{F}$, we can then add a notion of *probability measure* — as the name implies, this measures how likely an event is to occur.

Definition 2. A probability measure $\mathbb{P}$ on a σ-algebra $\mathcal{F}$ is a function $\mathbb{P}\colon \mathcal{F} \to \mathbb{R}$ which is non-negative, countably additive, and assigns measure 1 to the sample space. In symbols:

[1] I think the reasons why we need a σ-algebra are beyond the scope of this course, and certainly far beyond the context of this recitation. But if you want to chat a bit about what kind of problems arise without them, feel free to email me or stop by my office hours.

- For any $A \in \mathcal{F}$, we have $\mathbb{P}(A) \geq 0$ (**Non-negative**)
- For any $A_1, A_2, \cdots \in \mathcal{F}$ which are disjoint, we have $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$ (**Countably additive**)
- We have $\mathbb{P}(\Omega) = 1$ (**Assigns measure 1 to Ω**)

Together, the tuple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*.

The entirety of the calculus of probabilities follow from these axioms. A few important results are listed below — proofs can be found in the textbook.

Theorem 1 (Properties of probability measures)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $A_1, A_2 \in \mathcal{F}$. Then*

- *We have* $0 \leq \mathbb{P}(A) \leq 1$
- *We have* $\mathbb{P}(A^{\mathsf{c}}) = 1 - \mathbb{P}(A)$
- *We have* $\mathbb{P}(\emptyset) = 0$
- *If* $A_1 \subset A_2$*, then* $\mathbb{P}(A_1) \leq \mathbb{P}(A_2)$
- *We have* $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$

Under a frequentist view of probability, we can consider the probability to be the fraction of favorable outcomes of an event when an experiment is repeated. For example, we could interpret a probability of 0.45 for heads in a coin flip to mean that the coin would, on average, land on heads 45 times if flipped 100 times.

Remark
The mathematical theory of probability all rests on this measure-theoretic treatment. Our theory of probability generally will not. However, if you find this sort of thing cool, there are many great textbooks on rigorous probability theory, for example the books by Shiryaev [3] and Durrett [2]. It also lays foundations for talking about some harder things, like the theory of stochastic processes or filtering theory.

1.2 Conditional Probability

Often, we are interested in the probability of one event *conditioned* on another — i.e., the probability of some event A, given the knowledge that some event B occurs. The common example is that of a dice roll: assuming a fair dice, the probability of a 2 is $\frac{1}{6}$. However, if we know the result of the dice roll was even, then the probability is $\frac{1}{3}$. This intuition is captured in the following definition:

Definition 3. Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and events A, B, the *conditional probability of A given B*, denoted $\mathbb{P}(A \mid B)$, is defined by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

One useful aspect of conditional probabilities is the ability to decompose the probability of an event by using some conditional probabilities of that event. In order to state this theorem, we define a partition:

Definition 4. A *partition* of a sample space Ω is a collection of disjoint events $B_1, \ldots, B_n$ which *cover* Ω and are pairwise disjoint. In symbols:

- We have $\Omega \subset B_1 \cup \cdots \cup B_n$ (**The set of B_ks cover Ω**)
- We have $B_1 \cap \cdots \cap B_n = \emptyset$ (**Pairwise disjoint**)

Then we can state the following theorem.

Theorem 2 (Total Probability Theorem). *For an event A and partition $B_1, \ldots, B_n$ of Ω, we have*

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i). \tag{2}$$

We can define one of the most important concepts in probability based off of conditional probabilities: *independence.*

Definition 5. Two events A and B are said to be independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Alternatively, by applying Equation 2, we can write Independence as $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, i.e. knowledge of B occurring does not change our knowledge of A occurring.

1.3 Bayes' Theorem

From Equation 1, we can write

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A).$$

This lets us write down *Bayes' Theorem*.

Theorem 3 (Bayes' Theorem)**.** *For events A and B, we have*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{3}$$

Together, Bayes' Theorem and the Total Probability Theorem allow us to calculate some probabilities in terms of some others. This sounds vague and pointless, but is actually extremely useful in real life. We now work through some examples.

2 Examples

I worked through a number of examples for homework problems, both from the course textbook and other sources. To avoid cheating, they are not included in this version. If you took the course and want the original version with included problems, please reach out!

References

[1] C.G. Boncelet. *Probability, Statistics, and Random Signals*. The Oxford series in electrical and computer engineering. Oxford University Press, 2016. ISBN: 9780190200510.

[2] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.

[3] Albert N Shiryaev. *Probability-1*. Vol. 95. Springer, 2016.

Module 2 Recitation Notes

ESE 306
Summer Session II 2022
Dan Waxman

Abstract

We'll summarize the key concepts of the current module in a few pages, then thoroughly work through four or five problems. This material comes from the class slides or the class textbook [1] unless otherwise noted.

1 Key Concepts

1.1 Combinatorics

We begin with *combinatorics*, which is about counting combinations.

One of the first questions you may ask about combinations is how many *permutations*, or orderings, of n distinguishable objections there are. We can think of this constructively by choosing object 1, of which there are n choices, then choosing object 2, of which there are $n-1$, choices, and repeating until we've exhausted the number of objects.

This means the number of permutations of n distinguishable objects is

$$n! := n \times (n-1) \times \cdots \times 1. \tag{1}$$

Permutations are important, but sometimes we're only interested in permutations of some subset of objects. For example, perhaps we're interested in the number of 3-letter ”words.”

In order to achieve this, we begin with the total number of permutations, then divide it by the number of ways to arrange the remaining $n-k$ objects. Then we make the same construction and arrive at

$${}^{n}_{k}P := \frac{n!}{(n-k)!}. \tag{2}$$

Other times (actually, more commonly,) we're interested in the number of *combinations* of k objects out of n. The distinction here is that we no longer care about ordering: under permutations, we consider ”abc” and ”cba” as two distinct words, but in terms of combinations, they are the same.

Then we can arrive at the appropriate quantity by ”undoing” the number of permutations in ${}^n_kP$, i.e.

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{{}^n_kP}{k!}. \tag{3}$$

We can generalize this into *partitioning*, where we separate n objects into m classes of size $k_1, \ldots, k_m$.

The expression for the number of possible partitions with these quantities is given by

$$\frac{n!}{k_1! \cdot \cdots \cdot k_m!}. \tag{4}$$

Note that we can consider $\binom{n}{k}$ as the $m = 2$ special case here, with the partitions being ”chosen” or ”not chosen.”

One useful expression is a recursive relation for $\binom{n}{k}$:

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}. \tag{5}$$

Another really useful result in math is the Binomial Theorem, which allows us to express multinomial coefficients in terms of $\binom{n}{k}$.

Theorem 1 (Binomial Theorem)**.** *For any integer $n > 0$, we have*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}. \tag{6}$$

The proof for this is by induction using (5).

If we're concerned with the probability of choose k_1 elements from a group of n_1 objects, k_2 elements from a group of n_2 objects, up to k_r elements from a group of n_r objects, we arrive at a hypergeometric probability

$$\frac{\binom{n_1}{k_1}\binom{n_2}{k_2}\cdots\binom{n_r}{k_r}}{\binom{n}{k}}. \tag{7}$$

1.2 Discrete Random Variables

The definition of a random variable might not exactly be what you expect

Definition 1. A *random variable* (r.v.) is a (measurable[1]) function $X\colon \Omega \to \mathbb{R}$. If the image $\{X(\omega) | \omega \in \Omega\}$ is countable, then X is said to be a *discrete* r.v. If the image is uncountable, we say X is a *continuous* r.v. This image is called the *support* of the r.v.

Lots of the theoretical aspects of r.v.s are much easier to develop for the discrete case, so we start with those. Much of the discussion, however, will be similar to that of continuous r.v.s, mostly swapping out sums/differences for integrals/derivatives, modulo some measure-theoretic details.

The empirical values of a random variable X are described by its *cumulative distribution function.*

Definition 2. Given some random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$, we can define the cumulative distribution function (cdf) of X by

$$F_X(x) = \mathbb{P}(X \leq x). \tag{8}$$

Some important properties of the cdf are summarized below.

Theorem 2. *For discrete r.v. X:*

- $\lim_{x \to -\infty} F_X(x) = 0$*;*
- $\lim_{x \to +\infty} F_X(x) = 1$*;*
- $F_X(x)$ *is non-decreasing;*
- $F_X(x)$ *is right-continuous;*
- $\mathbb{P}(X > x) = 1 - F_X(x)$*;*
- $\mathbb{P}(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$*;*
- $\mathbb{P}(X = x) = F_X(x) - F_X(x^-)$*.*

Often, it's rather inconvenient to specify the cdf F_X, and is instead easier to specify the *probability mass function.*

Definition 3. Given some discrete r.v. X, the probability mass function (pmf) is the function defined by

$$f_X(x) \coloneqq P(X = x) = F_X(x) - F_X(x^-). \tag{9}$$

We immediately notice that for r.v. X with support $x_1, x_2, \ldots$, by the definition of probability,

$$\sum_{i=1}^{\infty} p(x_i) = 1. \tag{10}$$

The simplest r.v. is a *Bernoulli* r.v., whose pmf is described by a parameter p:

$$f_X(X = 1) = 1; \quad f(X = 0) = 1 - p. \tag{11}$$

We usually call $X = 1$ a *success* and $X = 0$ a *failure.*

If we take n independent and identically distributed Bernoulli trials, then the number of successes is a *binomial random variable.* If the Bernoulli trials have parameter p, then the pmf of a binomial r.v. is given by

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{12}$$

If we are instead concerned with the number of trials required for a success, then we have a *geometric random variable*, whose pmf[2] is given by

$$f_X(x) = (1-p)^{x-1} p. \tag{13}$$

Moving away from Bernoulli trials, we have another important r.v.: the *Poisson random variable.* Poisson r.v.s are often used to model the number of occurences of a rare event in some fixed time frame. Its pmf is given by

$$f_X(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \tag{14}$$

where $\lambda > 0$ is the sole parameter.

2 Examples

I worked through a number of examples for homework problems, both from the course textbook and other sources. To avoid cheating, they are not included in this version. If you took the course and want the original version with included problems, please reach out!

References

[1] C.G. Boncelet. *Probability, Statistics, and Random Signals.* The Oxford series in electrical and computer engineering. Oxford University Press, 2016. ISBN: 9780190200510.

[2] Take as warning that this is convention, and in some other texts the geometric r.v. is used to describe the number of trials failed before a success

Module 3 Recitation Notes

ESE 306
Summer Session II 2022
Dan Waxman

Abstract

We'll summarize the key concepts of the current module in a few pages, then thoroughly work through a few problems. This material comes from the class slides or the class textbook [1] unless otherwise noted.

1 Key Concepts

1.1 Review of Random Variables

We begin by reviewing the definition, and some basic properties, of (discrete or continuous) random variables.

Definition 1. A *random variable* (r.v.) is a (measurable[1]) function $X\colon \Omega \to \mathbb{R}$. If the image $\{X(\omega)|\omega \in \Omega\}$ is countable, then X is said to be a *discrete* r.v. If the image is uncountable, we say X is a *continuous* r.v. This image is called the *support* of the r.v.

The empirical values of a random variable X are described by its *cumulative distribution function*.

Definition 2. Given some random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$, we can define the cumulative distribution function (cdf) of X by

$$F_X(x) = \mathbb{P}(X \leq x). \tag{1}$$

Some important properties of the cdf are summarized below.

Theorem 1. *For r.v. X with cdf $F_X(x)$:*

- $\lim_{x\to-\infty} F_X(x) = 0$*;*
- $\lim_{x\to+\infty} F_X(x) = 1$*;*
- $F_X(x)$ *is non-decreasing;*

- *$F_X(x)$ is right-continuous;*
- *$\mathbb{P}(X > x) = 1 - F_X(x)$;*
- *$\mathbb{P}(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$;*
- *$\mathbb{P}(X = x) = F_X(x) - F_X(x^-)$.*

1.2 Continuous Random Variables

With this discussion of r.v.s/cdfs complete, we can turn to some properties of *continuous* random variables. Like mentioned last time, the measure theory gets much more difficult, but from our perspective the treatment will mostly just swap derivatives for differences and integrals for sums.

Just as it was difficult to deal with the cdf of a discrete r.v., it can be a pain (actually, even more so) to deal with the cdf of a continous r.v. However, describing a probability mass does not make sense, since for most continuous distributions $\mathbb{P}(X = x) = 0$ for any x.

So instead we take a slightly different approach: recall that we could define the pmf by taking the difference $F_X(x) - F_X(x^-)$. Instead of taking the difference, we could imagine taking the derivative, defining a *probability density function.* Let's state this a little differently.

Definition 3. For a continuous r.v. X with cdf $F_X(x)$, the probability density function (pdf) $f_X(x)$ is defined to satisfy

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du. \tag{2}$$

If $f_X(x)$ exists, we say that X *admits* the density f_X.

By taking the derivative of both sides of (2), we arrive at what was stated before, that

$$f_X(x) = \frac{dF_X}{dx}\bigg|_{x=x}.$$

Directly from properties of the cdf, we can derive a few properties of the pdf.

Theorem 2. *For r.v. X with pdf $f_X(x)$:*

- $\int_{-\infty}^{\infty} f_X(x) = 1$

- $\mathbb{P}(X = x) = 0$
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(u)\, du$

1.3 Some Important Continuous Distribution

We begin with perhaps the simplest distribution, the *uniform distribution.*

Definition 4. For $a < b$, an r.v. X is said to be distributed according to the uniform distribution, denoted $X \sim \mathcal{U}(a, b)$, if it admits pdf

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

In this case, the cdf is easy to recover:

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b. \end{cases}$$

Next, we consider the slightly more sophisticated, but still quite simple *exponential distirbution.*

Definition 5. For $\lambda > 0$, an r.v. X is said to be distributed according to the exponential distribution, denoted $X \sim \mathcal{E}(\lambda)$, if it admits pdf

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{4}$$

Again, calculating the cdf is not too difficult here:

$$F_X(x) = \int_0^x \lambda e^{-\lambda u}\, du = 1 - e^{-\lambda x}.$$

Finally, we finish with likely the most widely used continuous distribution, the *Gaussian (or Normal) distribution.*

Definition 6. For real μ and $\sigma^2 > 0$, an r.v. X is said to be normally distribution, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if it admits pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}(x-\mu)^2\right]. \tag{5}$$

For the Gaussian distribution, the parameter μ is called the *mean*, and the parameter σ^2 is called the *variance*. The square root of the variance, σ, is called the *standard deviation*. When $\mu = 0$ and $\sigma^2 = 1$, the resulting distribution $\mathcal{N}(0,1)$ is called the *standard normal*.

Juxtaposed to the previous few distributions we explored, it's actually not possible to write down the Gaussian cdf in terms of elementary functions. Instead, we use look-up tables (or, more realistically nowadays, a program) to calculate approximate values. We often write the cdf of the standard normal as $\Phi(x)$. Then by taking the r.v.

$$Z = \frac{X-\mu}{\sigma}, \tag{6}$$

which is normally distributed, we can use $\Phi(x)$ to calculate quantiles for any normal r.v.

2 Examples

I worked through a number of examples for homework problems, both from the course textbook and other sources. To avoid cheating, they are not included in this version. If you took the course and want the original version with included problems, please reach out!

References

[1] C.G. Boncelet. *Probability, Statistics, and Random Signals*. The Oxford series in electrical and computer engineering. Oxford University Press, 2016. ISBN: 9780190200510.

Module 4 Recitation Notes

ESE 306
Summer Session II 2022

Dan Waxman

Abstract

We'll summarize the key concepts of the current module in a few pages, then thoroughly work through a few problems. In particular, we will focus on joint random variables, as expectations are not on Quiz 4. This material comes from the class slides or the class textbook [1] unless otherwise noted.

1 Key Concepts

1.1 Joint Random Variables

Consider the case where we have a number of random variables, $X_1, X_2, \ldots, X_n$, which are related to each other. For example, we can take $n = 2$, and consider X_1 to be the height of a person and X_2 to be the weight of a person. Clearly, X_1 and X_2 are related: a taller person, on average, tends to be heavier.

It might make sense to model the height of people and weight of people separately, but since they're related it makes more sense to make a *joint* model; i.e., $p_{X_1,X_2}(x_1, x_2)$ giving the probability of some height X_1 *and* weight X_2. In this two-variable case, p_{X_1,X_2} is called a *bivariate distribution.*

Just as in the univariate case, there are a few ways to describe the empirical properties of joint random variables. We start with the CDF, which is the same in both the discrete and continuous case.

Definition 1. The *cumulative distribution function* of random variables $X_1, X_2, \ldots, X_n$ is a function $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) \colon \Omega_1 \times \cdots \times \Omega_n \to [0, 1]$ given by

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n). \tag{1}$$

The CDF has the properties one would expect from the univariate case. We list a few here.

Theorem 1. *Let X_1, X_2 be random variables with joint cdf $F_{X_1,X_2}(x_1, x_2)$ and marginal cdfs $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$. Then:*

- $F_{X_1}(x_1) = \lim_{x_2\to\infty} F_{X_1,X_2}(x_1, x_2)$.

- $F_{X_2}(x_2) = \lim_{x_1 \to \infty} F_{X_1,X_2}(x_1, x_2)$.
- $\lim_{x_1 \to -\infty} F_{X_1,X_2}(x_1, x_2) = 0$.
- $\lim_{x_2 \to -\infty} F_{X_1,X_2}(x_1, x_2) = 0$.
- *For* $x_1 \leq x_1'$ *and* $x_2 \leq x_2'$, *we have* $F_{X_1,X_2}(x_1, x_2) \leq F_{X_1,X_2}(x_1', x_2')$.
- $\lim_{(x_1,x_2) \to (\infty,\infty)} F_{X_1,X_2}(x_1, x_2) = 1$.

For discrete joint random variables, we use a pmf, and for continuous joint random variables, we use a pdf.

Definition 2. The *joint probability mass function* of discrete random variables $X_1, X_2, \dots, X_N$ is given by

$$p_{X_1,\dots,X_N}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_N = x_N). \tag{2}$$

The *joint probability density function* of continuous random variables $X_1, \dots, X_N$ with joint cdf $F_{X_1,\dots,X_N}$ is a function $f_{X_1,\dots,X_N}$ such that

$$F_{X_1,\dots,X_N}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} f_{X_1,\dots,X_N}(x_1, \dots, x_N)\, dx_1 \dots dx_N. \tag{3}$$

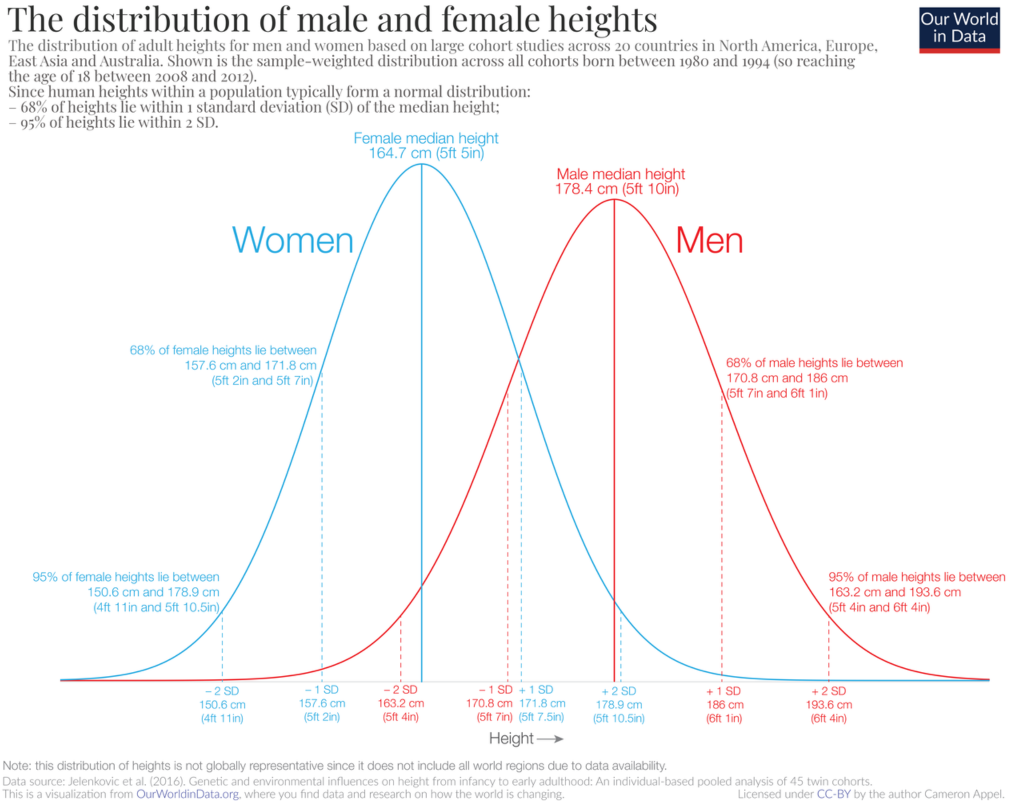
Note that the sum over all values of X_k for each k must be 1 in a set of discrete r.v.s, and the corresponding integral evaluate to unity in a set of continuous r.v.s. This is because the sum (integral) of a pmf (pdf) over a region B equals the probability that $(x_1, \dots, x_N) \in B$.

1.2 Marginal Probabilities

Recall the bivariate example of before, where X_1 is height and X_2 is weight. Despite these obviously being related to each other, it often makes sense to talk about the distribution of height, or the distribution of weight, separately. For example, consider the distribution of height for adult men gathered from data in North America, Europe, East Asia, and Australia in Figure 1 [2]. This can still be interesting and valuable information, despite not knowing specific weights, socioeconomic standing, etc. which also tend to influence this.

These sorts of distributions are called *marginal distributions*, and the process of getting $f_{X_1}(x_1)$ from $f_{X_1,X_2}(x_1, x_2)$ is called *marginalizing.* This looks similar for discrete and continuous random variables, with the different being a sum vs. an integral. For example, if X_1, X_2 were discrete, we'd write

$$p_{X_1}(x_1) = \sum_{x_2 \in \Omega_2} p_{X_1,X_2}(x_1, x_2).$$



The distribution of male and female heights

The distribution of adult heights for men and women based on large cohort studies across 20 countries in North America, Europe, East Asia and Australia. Shown is the sample-weighted distribution across all cohorts born between 1980 and 1994 (so reaching the age of 18 between 2008 and 2012).
Since human heights within a population typically form a normal distribution:
– 68% of heights lie within 1 standard deviation (SD) of the median height;
– 95% of heights lie within 2 SD.

Our World in Data
Female median height
164.7 cm (5ft 5in)
Male median height
178.4 cm (5ft 10in)
Women
Men
68% of female heights lie between
157.6 cm and 171.8 cm
(5ft 2in and 5ft 7in)
68% of male heights lie between
170.8 cm and 186 cm
(5ft 7in and 6ft 1in)
95% of female heights lie between
150.6 cm and 178.9 cm
(4ft 11in and 5ft 10.5in)
95% of male heights lie between
163.2 cm and 193.6 cm
(5ft 4in and 6ft 4in)
– 2 SD 150.6 cm (4ft 11in)
– 1 SD 157.6 cm (5ft 2in)
– 2 SD 163.2 cm (5ft 4in)
– 1 SD 170.8 cm (5ft 7in)
+ 1 SD 171.8 cm (5ft 7.5in)
+ 2 SD 178.9 cm (5ft 10.5in)
+ 1 SD 186 cm (6ft 1in)
+ 2 SD 193.6 cm (6ft 4in)
Height ⟶

Note: this distribution of heights is not globally representative since it does not include all world regions due to data availability.
Data source: Jelenkovic et al. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts.
This is a visualization from OurWorldinData.org, where you find data and research on how the world is changing. Licensed under CC-BY by the author Cameron Appel.

Figure 1: Distribution of male and female heights [2].

Meanwhile, if there were continuous, we'd write

$$f_{X_1}(x_1) = \int_{x_2 \in \Omega_2} f_{X_1,X_2}(x_1, x_2)\, dx_2.$$

Note that the marginal distribution f_{X_1} is different, but related to, the conditional distribution $f_{X_1|X_2}$. In particular, we can view the marginal distribution as a *weighted average* of the conditional distribution. This is because

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1|X_2}(x_1 \mid x_2) f_{X_2}(x_2), \tag{4}$$

so (e.g. in the continuous case)

$$\begin{aligned} f_{X_1}(x_1) &= \int_{\Omega_2} f_{X_1,X_2}(x_1, x_2)\, dx_2 \\ &= \int_{\Omega_2} f_{X_1|X_2}(x_1 \mid x_2) f_{X_2}(x_2)\, dx_2 \\ &= \mathbb{E}_{f_{X_2}}[f(x_1 \mid x_2)]. \end{aligned}$$

2 Examples

I worked through a number of examples for homework problems, both from the course textbook and other sources. To avoid cheating, they are not included in this version. If you took the course and want the original version with included problems, please reach out!

References

[1] C.G. Boncelet. *Probability, Statistics, and Random Signals.* The Oxford series in electrical and computer engineering. Oxford University Press, 2016. ISBN: 9780190200510.

[2] Cameron Appel Max Roser and Hannah Ritchie. "Human Height". In: *Our World in Data* (2013). https://ourworldindata.org/human-height.