University of California, Berkeley

Department of Industrial Engineering and Operations Research

John-Michael Laurel

Expanded lecture notes adapted from Anil Aswani's spring 2020 lecture compilation for IEOR-165 course on engineering statistics, quality control, and forecasting.

LECTURE NOTES

Contents

List of Figures

LECTURE 1

Introduction and Review of Probability

This course in segmented into two parts: *modeling and regression* and *hypothesis testing* with the latter being more difficult; which can be thought of as decision making. Regarding the courses contribution to the curriculum at Berkeley, especially with the data-driven courses (e.g. Data 8 and IEOR-142), nominally there will be overlap between those courses with this one. The hope for the end of the course is that one acquires a different philosophical approach in statistics. Aswani mentions that such a statement might sound odd at first, however what makes statistics as a discipline difficult is that there is a philosophy behind it. For example, in the state of analysis there are different methods to analyze the same data set and potentially end up with completely different results. Unless you have an understanding of the philosophy of how your conducting the analysis, it makes it difficult in designing a plan moving forward, but is also makes it difficult to evaluate the analysis that someone else has done.

For instance, court cases involving some statistical analysis. As one might expect, the statisticians from the different parties will reach completely different conclusions on the same data sets. Everything they do is completed with technical correctness, in the sense that they're running tests correctly, performing regression correctly, etc. So the question that arises again is "How in the world do two people performing the analysis on the same data get different results?" One of the things that can happen is in fact, one can begin manipulating their analysis to reach their desired outcome. This makes sense in the context of court cases, where statisticians are hired to produce a particular outcome changing the analysis in certain ways to reach particular outcome. Unless you have an appreciation of the philosophical approach, it's really difficult to detect whether manipulation of the analysis is in play.

Aswani emphasizes that different areas have different philosophical approaches. For instance, one that'll be referred to often for contrast is that economists have a completely different approach to analysis compared to statisticians. Ditto for data and computer scientists. The School of Public Health, School of Life Sciences, etc. have completely different philosophical approaches. Future discussion will include some introduction to these different philosophical approaches. For even more context, computer scientists really think about statistics very computationally, a mindset where analysis is an impersonal objective technical entity, taking data, pushing it into a computer for a formal calculation, and that result is correct. In contrast, economists believe they can determine cause and effect, not just correlation from analysis. Hence the phrase "correlation is not causation."

John-Michael Laurel. Tuesday, January 21, 2020

Statisticians support such a phrase, whilst economists will refute such a claim. In other words, correlation is suggestive of causation, but not conclusive. Medical professionals lean on this view, with the some going further in saying that the only way to get causation is through clinical trials.

In this course, we'll lean on the traditional statisticians side of analysis, rather than the computer scientists approach or economists approach. Debate still exists and by no means is settled. Let's begin.

1.1 Probability as a Measure

To make the notion of probability rigorous, we construct a *probability measure space* as a tuple of three elements:

$$(\Omega, \mathcal{F}, \mathbb{P}) \text{ , where } \begin{cases} \Omega = \text{the outcome space of some experiment,} \\ \mathcal{F} = \text{set of events of the outcome space, and} \\ \mathbb{P} = \text{probability measure on the set of events.} \end{cases}$$

Formally $\mathcal{F}$ is a σ-field, that is, a) a collection of subsets of Ω that includes Ω itself, b) is closed under complements, and c) is closed under countable unions. Intuitively, a σ-field is a set possessing the ability to combine outcomes in a sensible way to produce an event. The measure $\mathbb{P}$ is a function of $\mathcal{F}$ satisfies the following properties:

1.1.1 Axioms of Probability

Axioms of Probability (Kolmogorov)

A probability measure $\mathbb{P}$, belonging to some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with Ω the outcome space and $\mathcal{F}$, the set of events satisfies the following:

(A1) $\mathbb{P}(A) \geqslant 0$, (non-negativity)

(A2) $\mathbb{P}(\Omega) = 1$, and (normalization)

(A3) $\mathbb{P}(A_1 \cup A_2 \cup \dots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \dots$ (countable additivity)

, where $A \subseteq \Omega$ and $A_1, A_2, \dots$ is any countable collection of mutually exclusive events.

In words, probabilities of events are non-negative, the measure of the entire outcome space is 1, and the union of disjoint regions add. The first axiom is common sense, the second implies that something in the outcome space will happen, and the last axiom can be related to measures of disjoint regions e.g., measuring disjoint intervals on the real set of numbers is measuring each interval and taking their sum. It's prudent to note that an *outcome* is a member of Ω and an *event* is a member of $\mathcal{F}$. A concrete example will give texture to these axioms.

example 1.1 Toss a fair coin twice. Make explicit $(\Omega, \mathcal{F}, \mathbb{P})$.

solution 1.1 The outcome space is clearly

$$\Omega = \{HH, HT, TH, TT\}.$$

The σ-field $\mathcal{F}$ is given by the power set of Ω, meaning that

$$\mathcal{F} = \left\{ \begin{array}{cccc} \varnothing & \{HH\} & \{HT\} & \{TH\} \\ \{TT\} & \{HH, HT\} & \{HH, TH\} & \{HH, TT\} \\ \{HT, TH\} & \{HT, TT\} & \{TH, TT\} & \{HH, HT, TH\} \\ \{HH, HT, TT\} & \{HH, TH, TT\} & \{HT, TH, TT\} & \Omega \end{array} \right\}.$$

One possible measure $\mathbb{P}$ is given by the function defined as

$$\mathbb{P}(HH) = \mathbb{P}(HT) = \mathbb{P}(TH) = \mathbb{P}(TT) = \frac{1}{4}.$$

1.2 Consequences of the Kolmogorov's Axioms

Useful consequences arise from Kolmogorov's axioms; here are a few of them. The proofs are left as an exercise to the reader.

1.2.1 Equally Likely Outcomes

For an outcome space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ in which each outcome ω_j is equally likely, it holds for $1 \leqslant j \leqslant n$ that $\mathbb{P}(\omega_j) = 1/n$.

Figure 1.1 Set diagram: equally likely outcomes. In this realization, each outcome ω_j is depicted by a square with Area(ω_j) =Area(ω_i). This makes sense for equally likely outcomes, since $\mathbb{P}(\omega_j) = 1/n$ for $1 \leqslant j \leqslant n$. Clearly, $n = 12$ for this diagram. It is not necessary that Ω be cut into squares, any shape works so long that Area(ω_j) =Area(ω_i).

1.2.2 Complement Rule

If A^c denotes the *complement* of some event A, then $\mathbb{P}\left(A^c\right) = 1 - \mathbb{P}\left(A\right)$.

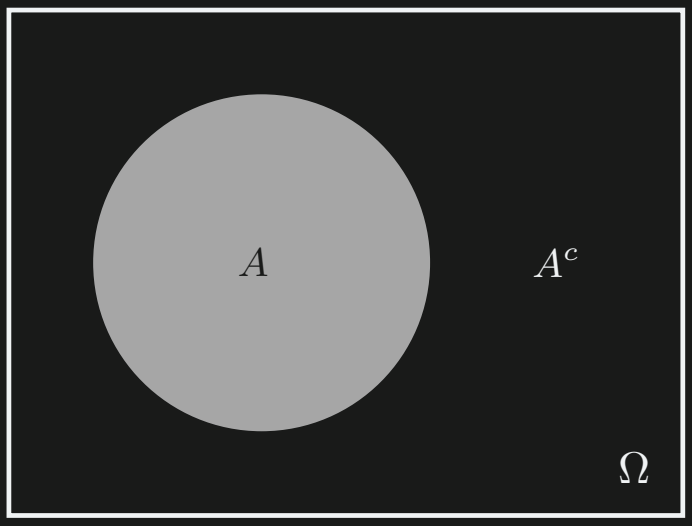

Figure 1.2 Complement rule. It's obvious that $A \cup A^c = \Omega$, hence $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$. In complex situations applying the rule of complements can mitigate unnecessary computations. e.g., the birthday problem.

1.2.3 Inclusion-Exclusion Principle

For two event A and B, not necessarily independent

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

Expanding this idea to n events, $A_1, A_2, \ldots, A_n$, not necessarily independent, we have

$$\mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} A_j - \sum_{i<j} \mathbb{P}(A_i A_j) + \cdots + (-1)^{n+1} \mathbb{P}(A_1 A_2 \cdots A_n).$$

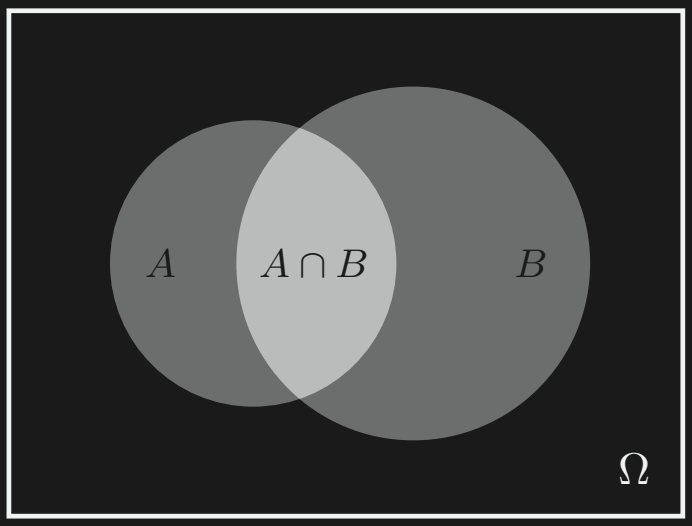

Figure 1.3 Inclusion-Exclusion Principle for two not necessarily independent events. Intuitively, taking $\mathbb{P}(A)$ and $\mathbb{P}(B)$ double counts their intersection $A \cap B$, hence we subtract out $\mathbb{P}(A \cap B)$. It's a good exercise to sort out how three not necessarily independent events would look like and extracting $\mathbb{P}(\cup_j A_j)$.

1.2.4 Sub-additivity

For two events $A \subseteq B$, $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.

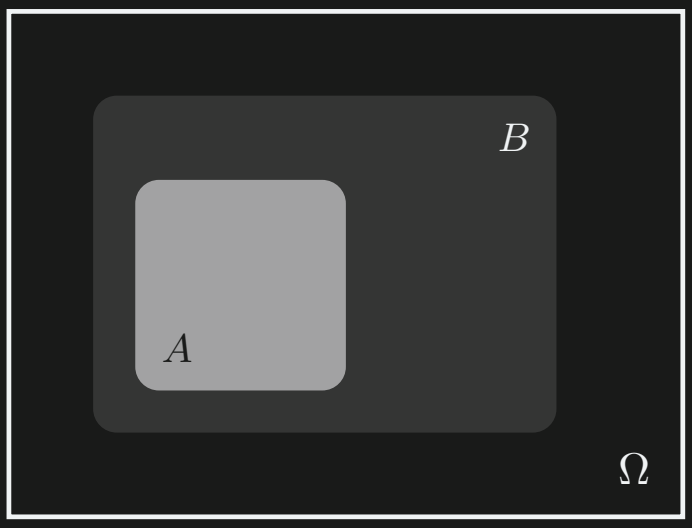

Figure 1.4 Set diagram: sub-additivity. Intuitively if $A \subseteq B$, then it's necessarily true that $\mathbb{P}(A) \leqslant \mathbb{P}(B)$. Obviously, $A \implies B$. i.e., event A implies event B.

1.2.5 Total Law of Probability

Consider a finite collection of mutually exclusive events $B_1, B_2, \ldots, B_n$ such that $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$ and $\mathbb{P}(B_j) > 0$. For any event A, it follows

$$\mathbb{P}(A) = \sum_{j=1}^{n} \mathbb{P}(B_j \cap A).$$

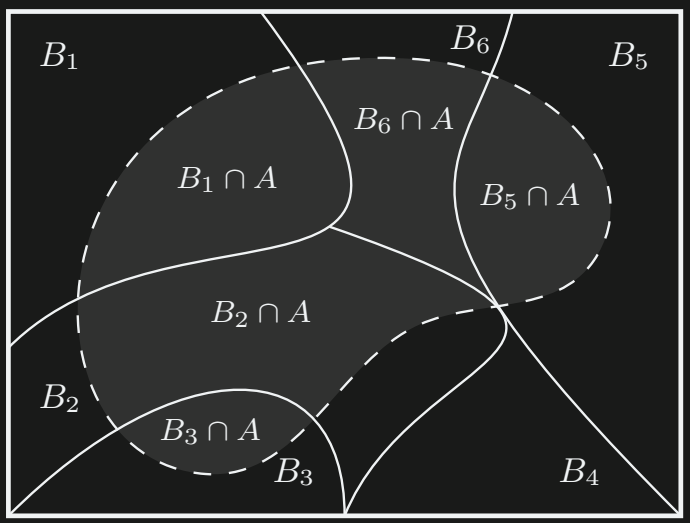

Figure 1.5 Total law of probability. In this realization, $B_1, B_2, \ldots, B_6$ form a complete partition of the outcome space Ω. The event A, depicted as a bean can be decomposed into the intersections of A and the B_j, then taking their sum. That is, $\mathbb{P}(A) = \sum_{j=1}^{6} \mathbb{P}(B_j \cap A)$, which is quite intuitive from the picture. It is not necessary that A touch every B_j. This is present in the diagram, where $B_4 \cap A = \varnothing$, which implies $\mathbb{P}(B_4 \cap A) = 0$.

1.3 Conditional Probability

Conditional probabilities arise from *dependent* events. We say the conditional probability of event B given A is defined as

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

Parsing the above, measure $B \cap A$ and re-normalize over the new universe A. Of course, it's stipulated that $\mathbb{P}(A) > 0$. From this, we can express the total law of probability as

1.3.1 Total Law of Probability

Total Law of Probability

For a finite collection of mutually exclusive events $B_1, B_2, \ldots, B_n$ such that $\Omega = \cup_{j=1}^{n} B_j$ and $\mathbb{P}(B_j) > 0$. For any event A, it follows

$$\mathbb{P}(A) = \sum_{j=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A \mid B_j)$$

, since $\mathbb{P}(B_j \cap A) = \mathbb{P}(B_j)\mathbb{P}(A \mid B_j)$ for $1 \leqslant j \leqslant n$.

This is rather powerful, since the event A can be decomposed into a sum of conditional probabilities, where the conditioning event B_i can be thought of as the different cases contributing to A. It's emphasized that this construction arises naturally from the definitions. Stitching these ideas together, we have the very powerful *Bayes theorem*.

1.3.2 Bayes' Theorem

Bayes' Theorem

For a finite collection of mutually exclusive events $B_1, B_2, \ldots, B_n$ such that $\Omega = \cup_{j=1}^{m} B_j$ and $\mathbb{P}(B_j) > 0$,

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(B_j)\ \mathbb{P}(A \mid B_j)}{\mathbb{P}(B_1)\mathbb{P}(A \mid B_1) + \cdots + \mathbb{P}(B_n)\mathbb{P}(A \mid B_n)}.$$

It's prudent to observe that the numerator is one of the terms in the sum in the denominator. This is always the case. Stripping away the normalizing constant $\mathbb{P}(A)$ we have

$$\mathbb{P}(B_j \mid A) \propto \mathbb{P}(B_j)\mathbb{P}(A \mid B_j),$$

where $\mathbb{P}(B_j)$ is referred to as the *prior* and $\mathbb{P}(A \mid B_i)$ the *posterior* probabilities. See Bayesian inference entry in `Wikipedia`. You can think of Bayes' theorem as an *inferential* direction, in that we seek to 'see where we came from given what we have.' This is often contrasted to the *causal* direction, 'given where we are, where do we go?'

example 1.2 Flip a fair coin twice. Compute $\mathbb{P}(HT \mid \{HT, TH\})$

solution 1.2 Intuitively, we know this to be $\frac{1}{2}$. By the definition of conditional probability,

$$\begin{aligned}\mathbb{P}\big(HT \mid \{HT, TH\}\big) &= \frac{\mathbb{P}(HT \cap \{HT, TH\})}{\mathbb{P}(HT, TH)} \\ &= \frac{\mathbb{P}(HT)}{\mathbb{P}(HT) + \mathbb{P}(TH)} \\ &= \frac{1}{2}\end{aligned}$$

example 1.3 Flip a fair coin twice. Compute $\mathbb{P}(\{HT, TH\} \mid HT)$

solution 1.3 Intuitively, we know this to be 1. By Bayes' theorem,

$$\begin{aligned}&\mathbb{P}\big(\{HT, TH\} \mid HT\big) \\ &= \frac{\mathbb{P}\big(\{HT, TH\}\big)\ \mathbb{P}\big(HT \mid \{HT, TH\}\big)}{\mathbb{P}\big(\{HT, TH\}\big)\ \mathbb{P}\big(HT \mid \{HT, TH\}\big) + \mathbb{P}\big(\{HT, TH\}\big)\ \mathbb{P}\big(TH \mid \{HT, TH\}\big)} \\ &= 1\end{aligned}$$

1.4 Independence

Two events A and B are *independent* iFF $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Equivalently, A and B are independent iFF $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

1.4.1 Independence of Multiple Events

Independence of Multiple Events

Multiple events $A_1, A_2, \ldots, A_m$ are *mutually independent* iFF for every subset of events

$$\{A_{j_1}, A_{j_2}, \ldots, A_{j_m}\} \subseteq \{A_1, A_2, \ldots, A_m\}$$

, the following holds:

$$\mathbb{P}\left(\bigcap_{i=1}^{m} A_{j_i}\right) = \prod_{i=1}^{m} \mathbb{P}(A_{j_i})$$

In words, when the probability of the intersection of events (A_{j_i}) factors, the events are independent. Conceptually in this course, independence will be important, because we will be thinking about data

as being drawn from some probability distribution while assuming that each data point is mutually independent of every other data point. Obviously, this assumption can fail in real data sets. One should think carefully where data is coming from in order to understand what methods of analysis are applicable. Aswani emphasizes that it is far more important to understand what an output from a program really means, that is how one should interpret an output. The motion of loading a data set and calling a function to push out answer is trivial. The assumption of mutual independence might seem innocuous for a data set, particularly UNCLEAR.

An exemplar setting where the assumption of mutual independence is noxious is network data. Suppose you're looking at Facebook data from Berkeley students with the intention of performing analysis on their political beliefs and using such analysis to say something about the political beliefs of the United States as a whole. This is problematic due to the inherent friend structure on Facebook, implying relatedness between the data points; which in turn makes the assumption of mutual independence harmful in producing a reliable result. This issue is seen empirically in the analysis of social networks in general. Some of the methods we develop in this course will fail under the presence of such relatedness and fundamentally this lack of independence can potentially impede the ability to produce a reliable conclusion from analysis.

1.4.2 Pair-Wise Independence

Multiple events $A_1, A_2, \ldots, A_m$ are *pair-wise* independent iFF every pair of events is independent, meaning $\mathbb{P}(A_j \cap A_{j'}) = \mathbb{P}(A_j)\mathbb{P}(A_{j'})$ for all distinct pairs of indices (j, j'). Pair-wise independence does not always imply mutual independence. That is to say that mutual independence is stronger than pair-wise independence.

example 1.4 Toss a fair coin twice. Are $\{HH, HT\}$ and $\{HT, TT\}$ independent?

solution 1.4 Intuitively, yes since the event of getting heads on the first toss and tails on the second toss is independent. Applying the definitions,

$$\mathbb{P}\big(\{HH, HT\} \cap \{HT, TT\}\big) = \mathbb{P}(HT) = \frac{1}{4} = \mathbb{P}\big(\{HH, HT\}\big)\mathbb{P}\big(\{HT, TT\}\big)$$

In other words, the measure $\mathbb{P}$ is defined such that the result of the first coin flip does not impact the result of the second coin flip.

example 1.5 Toss a fair coin twice. Are $\{HH, HT\}$ and $\{TT\}$ independent?

solution 1.5 Intuitively, no since tossing tails twice implies one did not toss heads. On the other hand, the first toss being H makes it impossible to get to tails. Applying the definitions,

$$\mathbb{P}\big(\{HH, HT\} \cap \{TT\}\big) = \mathbb{P}(\varnothing) = 0 \neq \mathbb{P}\big(\{HH, HT\}\big)\mathbb{P}\big(\{TT\}\big)$$

Restated, observing $\{HH, HT\}$ provides information about $\mathbb{P}(\{TT\})$. In contrast, observing $\{HH, HT\}$ doesn't provide information about $\mathbb{P}(\{HT, TT\})$.

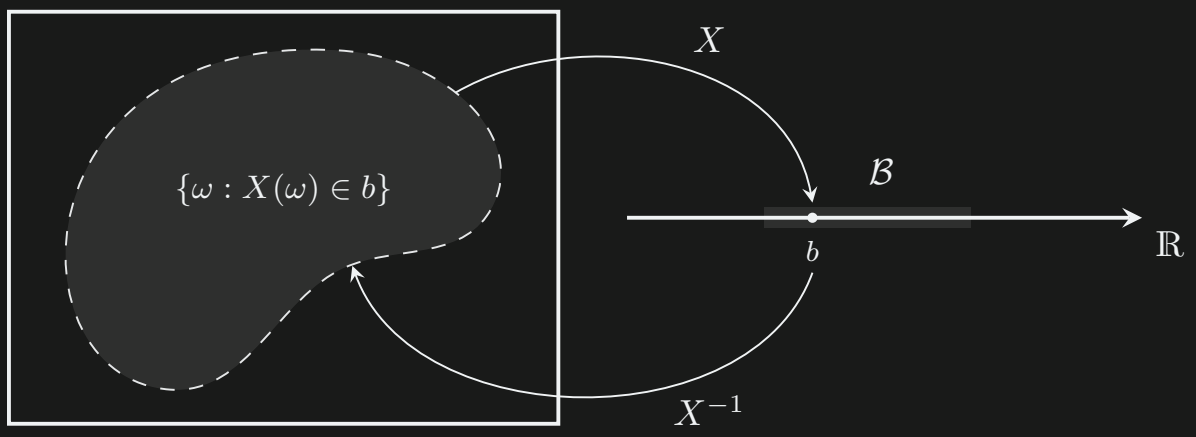

Figure 1.6 Random variable mapping. Here, $X^{-1}(b)$ is the pre-image of $b \in \mathcal{B}$ under X. i.e., a subset of Ω satisfying $\{\omega : X(\omega) \in b\}$. In the 'tossing a fair coin twice example,' $X^{-1}(1)$ corresponds to the event of tossing one head and is precisely $\{\omega : X(\omega) = 1\} = \{HT, TH\}$.

1.5 Random Variables

Formally, a random variable is a function $X : \Omega \to \mathcal{B}$ that maps the outcome space where Ω to a subset of the real numbers, $\mathcal{B} \subseteq \mathbb{R}$ with the property that the set $\{\omega : X(\omega) \in b\} = X^{-1}(b)$ is an event for every $b \in \mathcal{B}$. The formality of a random variable is illuminated with the above diagram.

1.6 The Cumulative Distribution Function

The cumulative distribution function (CDF) of a random variable X is defined by

Cumulative Distribution Function

The *cumulative distribution function* (CDF) of a random variable X is defined by

$$F_X(x) := \mathbb{P}\big(\{\omega : X(\omega) \leqslant x\}\big)$$

and completely characterizes the probability distribution of X over the real line.

The probability density function (PDF) of X is any function $f_X(x)$ such that

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx,$$

for any well-behaved set A. The units of $f_X(x)$ can be thought of as 'probability per unit length.' The PDF and CDF are closely related with the following equality

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

Because of this relationship, we assume that $F_X(x)$ is non-differentiable at only a finite number of points on its domain. Hence for CDFs that are mixtures of both discrete and continuous components, we have a special tool to 'force' a function to be continuous under integration. This is the topic of the next section.

1.7 Dirac-Delta Function, Brisk Overview

The *Dirac-Delta function*, denoted $\delta(\,\cdot\,)$ allows us to formulate the PDF of any discrete random variable. That's right, no need for the PMF. We start with the *Heaviside function*, which is defined as follows

$$H(t) = \begin{cases} 1, & t \geqslant 0 \\ 0, & \text{else} \end{cases}.$$

It's a very simple function, now differentiate H to get (something crazy)

$$\frac{dH(t)}{dt} = \begin{cases} +\infty & , t = 0 \\ 0 & , \text{ else} \end{cases}.$$

The Dirac-Delta function can be defined as

$$\delta(t) := \frac{dH(t)}{dt},$$

which by construction blows up to $+\infty$ when $t = 0$ and is 0 otherwise. This is not useful for us, however take a look $\delta(t)$ *under the integral*:

$$\int_{\mathbb{R}} \delta(t)dt = H(t)\Big|_{\mathbb{R}} = 1.$$

Magical right? Reason being that integrating over a function with a point mass of $+\infty$ at 0 and value 0 everywhere else pushes out one. If you're confused, realize this result follows immediately from $H(t)$. In a way, you can think of $\delta(t)$ as a switch that turns on to 1 if $t = 0$ *under the integral*. From this ideation, we can derive two very intuitive properties:

Shifting and Scaling Properties of $\delta(\,\cdot\,)$

For f, a real-valued function and T a constant, it is rather obvious:

(P1) $\displaystyle\int_{\mathbb{R}} \delta(t - T)dt = 1$ and (shifting invariance)

(P2) $\displaystyle\int_{\mathbb{R}} f(t)\delta(t)dt = f(0).$ (scaling by a function)

Stitching these two properties together, we have

$$\int_{\mathbb{R}} f(t)\delta(t - T)dt = f(T)$$

which is the mechanism through which we can formulate the probability density function for any discrete random variable.

Parsing the above, integrating $f(t)\delta(t-T)$ over $\mathbb{R}$ is equivalent to evaluating $f(t)$ at T. This makes perfect sense. Why? We're now entering a heuristic discussion for proving the scaling and shifting properties of $\delta(\,\cdot\,)$ under the integral. Treating $\delta(\,\cdot\,)$ as a switch under the integral means that scaling $\delta(\,\cdot\,)$ by a constant under the integral pushes out that constant. So for any function $f(t)$, the 'coefficient' of $\delta(t-T)$ under the integral pushes out that function evaluated at T, which is precisely $f(T)$.

□

You could also prove the combined scaling and shifting property of $\delta(\,\cdot\,)$ under the integral through integration by parts. Just assume that $f(t)$ vanishes as $t \to \infty$ for the integration to work. It's a completely valid argument to make since $\delta(\,\cdot\,)$ assumes the value zero everywhere except when **argmin** $\delta(\,\cdot\,) = +\infty$. Here's a start to the proof

Proof.

$$\int_{\mathbb{R}} f(t)\delta(t-T)dt = f(t)H(t-T)\Big|_{\mathbb{R}} - \int_{\mathbb{R}} f'(t)\delta(t-T)dt$$

$$\vdots$$

$$= f(T)$$

□

1.7.1 Descriptions of the Dirac-Delta Function

Dirac-Delta Function

The following descriptions of the Dirac-Delta function are *equivalent*:

- For a set A, the measure $\delta(A) = \begin{cases} 1, & 0 \in A \\ 0, & \text{else} \end{cases}$.
- The function $\delta(t)$ satisfies $\delta(t) = \begin{cases} +\infty & , t = 0 \\ 0 & , \text{else} \end{cases}$ and $\int_{\mathbb{R}} \delta(t)dt = 1.$

A concrete example of constructing a PDF using $\delta(\,\cdot\,)$ follows in the next example. The keen reader might have a slight objection to $\delta(\,\cdot\,)$, which is $\delta(\,\cdot\,)$ is not really a function, but nonetheless is a powerful construction for our purposes. From all this minutia, the next natural question is

> "What does $\delta(\,\cdot\,)$ actually look like?"

There are conventions for how to best visualize the Dirac-delta function, see the next page to get a feel for $H(\,\cdot\,)$ and $\delta(\,\cdot\,)$, then proceed to the next page to see what $f(t)\delta(t-T)$ looks like *under the integral*.

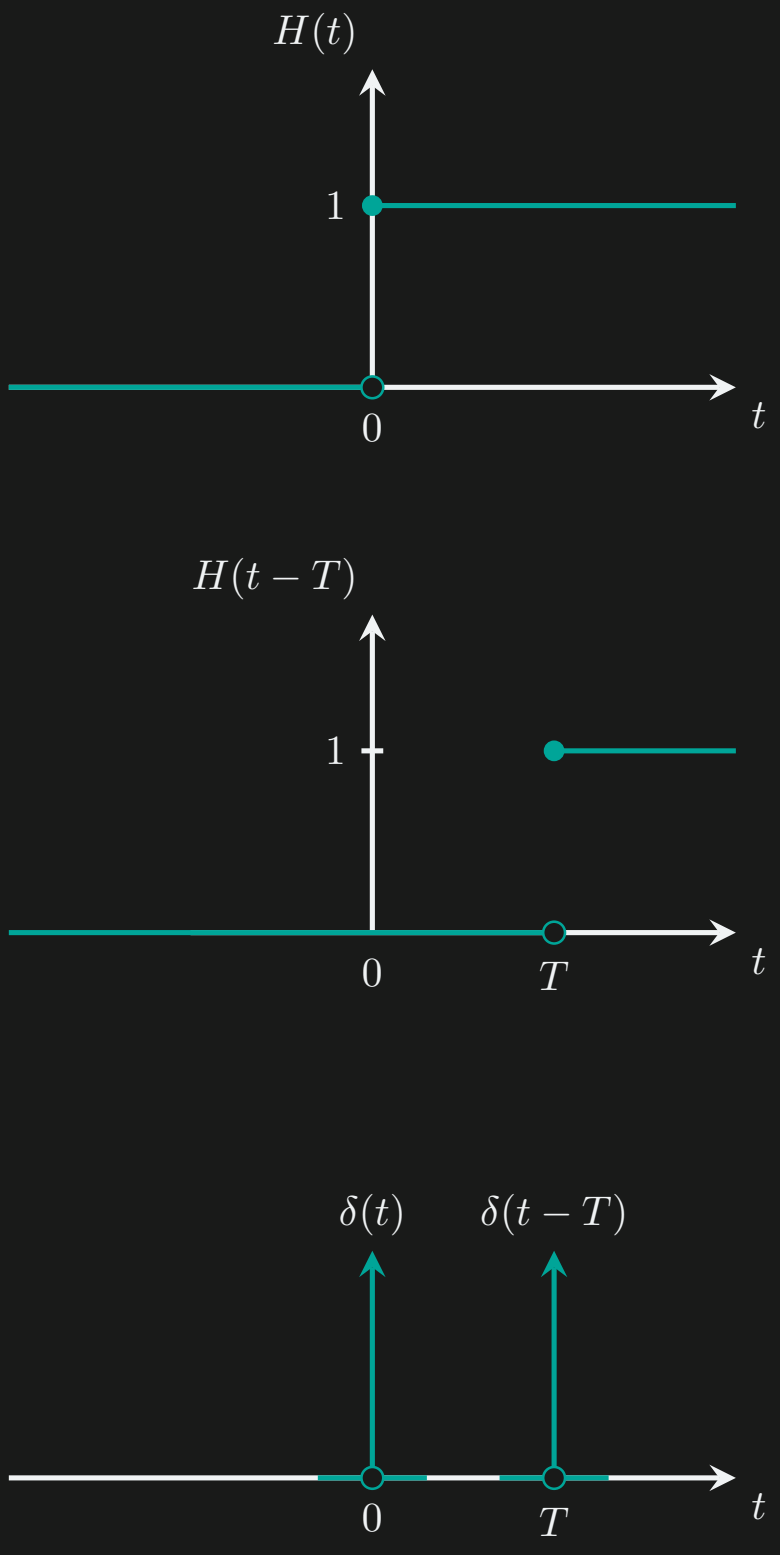

Figure 1.7 Heaviside function and its derivative: the Dirac-delta function, along with their shifted counterparts. The assigned value of infinity at $\{s : \textbf{argmin}\, \delta(s) = +\infty\}$ is depicted using a ray pointing upwards. For $\{s : \textbf{argmin}\, \delta(s) \neq +\infty\}$, $\delta(s)$ is assigned the value 0. This follows directly from $H'(s) = \delta(s)$.

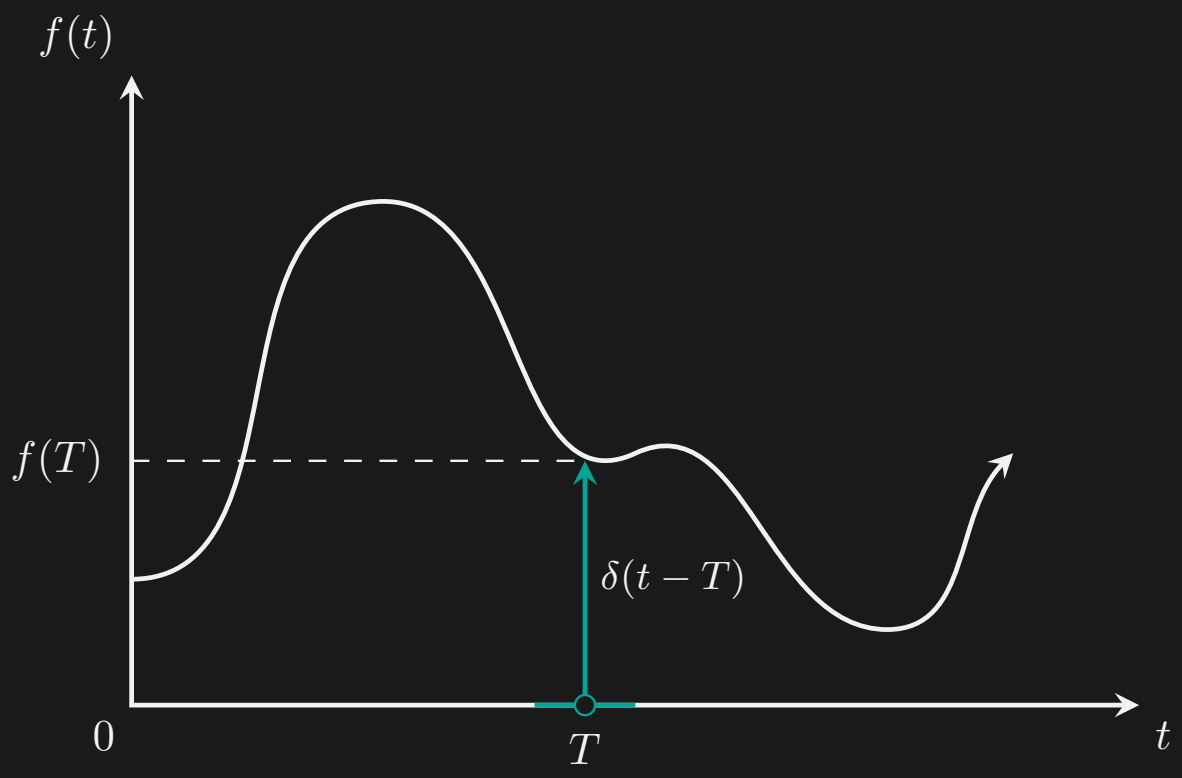

Figure 1.8 Dirac-delta function scaled by a function. When $\delta(t-T)$ is scaled by some function, say $f(t)$, the depiction of $\delta(\,\cdot\,)$ as a ray evolves into a vector. In particular, the length of the vector is precisely $f(T)$. This makes perfect sense. Intuitively, scaling $\delta(\,\cdot\,)$ by a constant simply scales $\delta(\,\cdot\,)$ by that constant under the integral. That is, $\int_{\mathbb{R}} c\delta(s)ds = c$ for $c \in \mathbb{R}$. Going back to the 'switch' analogy, *under the integral* $\delta(t-T)$ turns on at $t = T$ and is off everywhere $t \neq T$. Hence scaling $\delta(t-T)$ by $f(t)$ under the integral will assign the value $f(T)$ at $t = T$. This geometric interpretation coincides to the outline proof given earlier using integration parts. Again it's not necessary that $f(t)$ actually vanishes as $t \to \infty$. That assumption was the mechanism through which the integral worked out to our liking. From the depiction above, you should be convinced that it's not a stretch to make such an assumption and to not reap any harmful consequences.

example 1.6 Toss a fair coin twice and let X denote the number of heads. Construct its CDF and PDF.

solution 1.6 Accumulating probabilities, i.e., computing $\sum_{x=0}^{k} \mathbb{P}(X = x)$ for $k \in \{0, 1, 2\}$, the CDF is

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}, & 0 \leqslant x < 1 \\ \frac{3}{4}, & 1 \leqslant x < 2 \\ 1, & x \geqslant 2 \end{cases}$$

For the PDF, we utilize $\delta(\,\cdot\,)$ and it's easily seen that

$$f_X(x) = \frac{1}{4}\delta(x - 0) + \frac{1}{2}\delta(x - 1) + \frac{1}{4}\delta(x - 2).$$

This makes perfect sense under the integral. That is, integrating $f_X(x)$ over $\mathbb{R}$ produces 1, as required for a density. For clarity, using the definitions for $\delta(\,\cdot\,)$ it's obvious, for example that $\int_{\mathbb{R}} \frac{1}{2}\delta(x - 1) = \frac{1}{2}$. A plot of the $F_X(x)$ (above), $f_X(x)$, and the PMF of X helps in understanding this circle of ideas, see below.

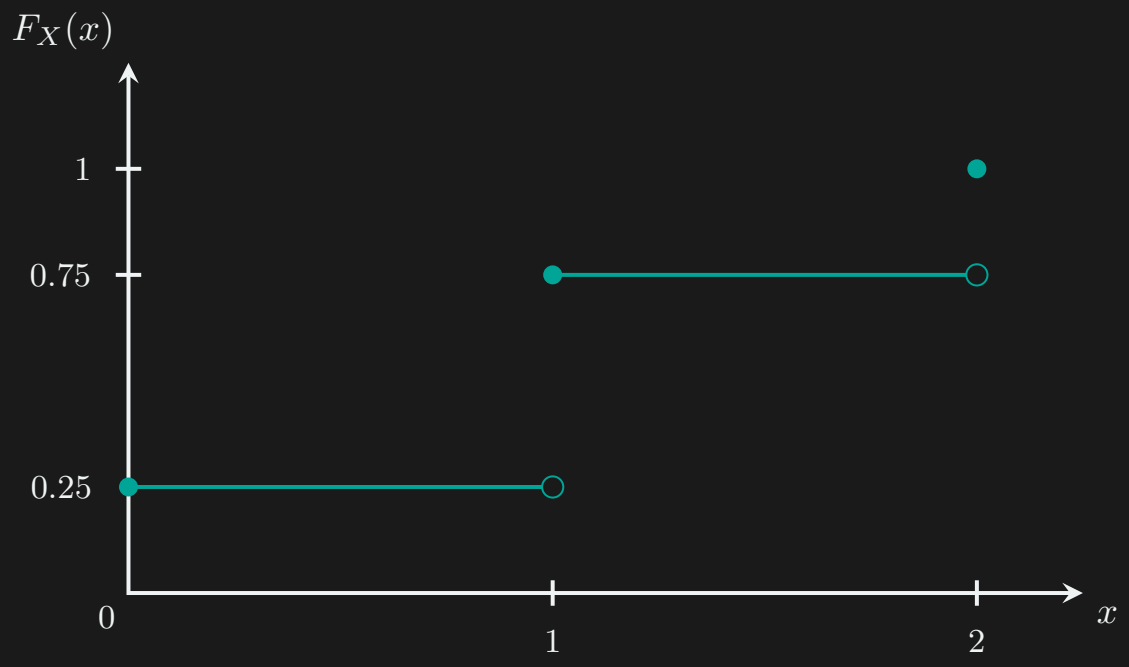

Figure 1.9 CDF for X = the number of heads in two tosses of a fair coin. It should be obvious why $\delta(\,\cdot\,)$ plays a critical role in constructing the PDF of X.

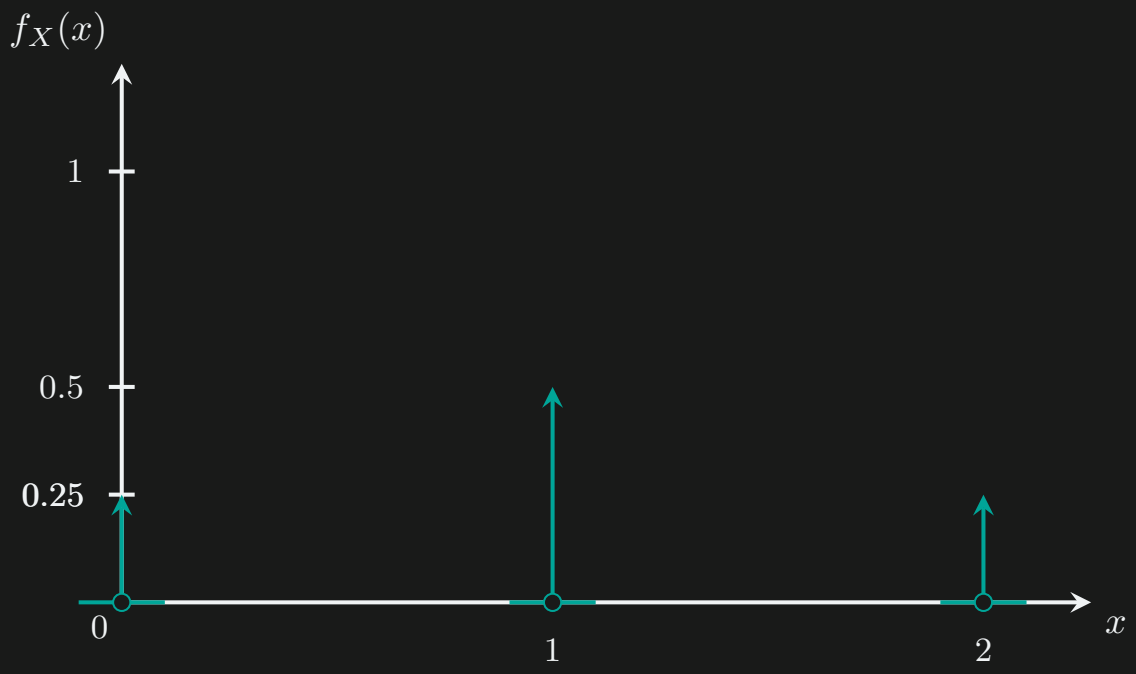

Figure 1.10 Probability density function: tossing a fair coin twice. Each upward facing vector represents $\delta(\cdot)$ scaled by some constant (the height of the vector) which is obviously $\mathbb{P}(X = x)$. Again integrating across $\mathbb{R}$ yields 1 as necessary. The Delta-dirac function makes sense under the integral.

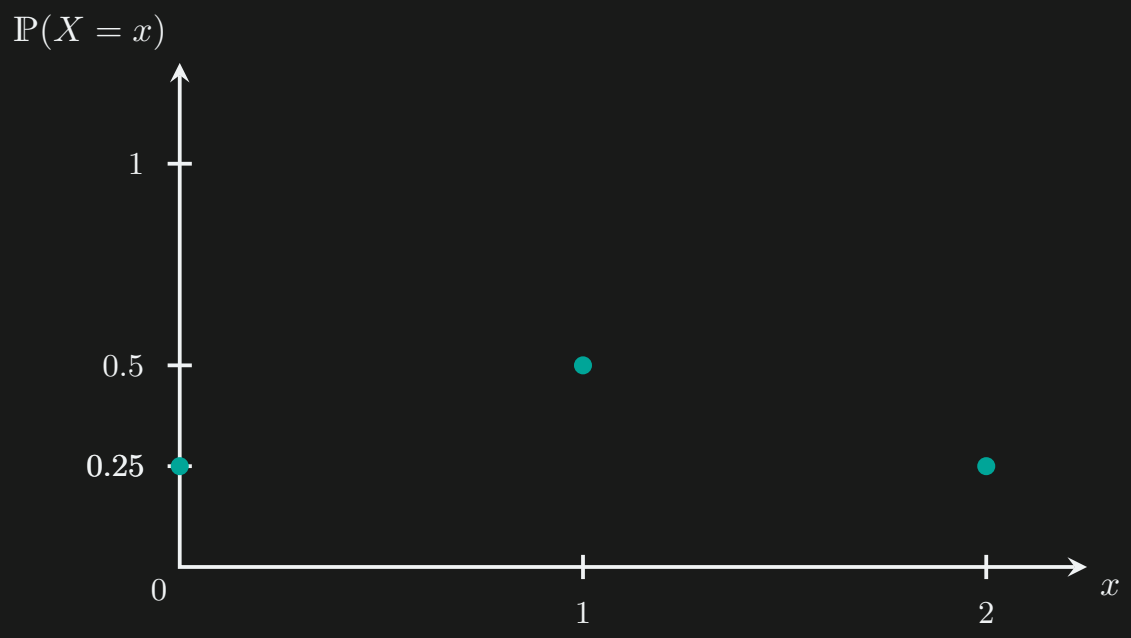

Figure 1.11 Probability mass function: tossing a fair coin twice. Sequence of point probabilities $(x, \mathbb{P}(X = x))$ for $X \in \{1, 2, 3\}$. Integrating here makes no sense, hence the need for $\delta(\cdot)$. Summing $\mathbb{P}(X = x)$ over $\{1, 2, 3\}$ does make sense.

1.8 Expectation

The *expectation* of a random variable X is the sum of the values it takes weighted by their probabilities. Formally,

Expectation of a Random Variable

For a random variable X and function $g : \mathbb{R} \to \mathbb{R}$, the expectation of $g(X)$ is

$$\mathbb{E}g(X) = \begin{cases} \sum_x g(x)\mathbb{P}(X = x) & \text{, for } X \text{ discrete} \\ \int_x g(x)\mathbb{P}(X \in dx) & \text{, for } X \text{ continuous.} \end{cases}$$

Of course, this idea extends to multivariate functions. i.e., one can substitute $h : \mathbb{R}^d \to \mathbb{R}$ for $g(X)$ and obtain a natural extension of the above constructions. Two important cases of expectation relate to statistical measures of a distribution, namely the mean and variance of a random variable.

1.8.1 Mean and Variance

Mean and Variance of a Random Variable

For a random variable X, the *mean* and *variance* of X are defined respectively

$$\mu(X) := \mathbb{E}X \quad \text{and} \quad \sigma^2(X) := \mathbb{E}(X - \mu)^2$$

The '(X)' will often be dropped for notational simplicity, being clear from context. Also, $\sigma^2(X) = \mathbb{V}\mathrm{ar}(X)$ will be used interchangeably. We think of $\mathbb{V}\mathrm{ar}(X)$ as the 'spread' or 'dispersion' of a distribution from its mean. Parsing the definition, it is the average squared distance of X from its mean. A computational form arises for variance:

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}X^2 - \mathbb{E}^2X\,.$$

In words, the variance of random variable is the expectation of the random variable squared minus the square of its expectation. The computational form follows from its definition. The *Law of Iterated Expectation*[1] is very intuitive and can help with computing expectation.

Law of Iterated Expectation

For random variables X and Y,

$$\mathbb{E}X = \mathbb{E}\big[\mathbb{E}(X \mid Y)\big].$$

The above is always true for *any* random variable. It's just a matter to figure out what random variable to condition on. A cute problem is showing $\mathbb{E}X_n = \frac{S_n}{n}$ for $S_n := \sum_{j=1}^n X_j$.

[1] Sometimes called the *towering rule* for whatever reason.

1.8.2 Properties of Mean and Variance

Properties of Mean and Variance

For $X_1, X_2, \ldots, X_n$ a sequence of random variables and $\alpha_j, \beta_j \in \mathbb{R}$, we have the following important properties:

- Expectation is *linear*.

$$\mathbb{E}\left[\sum_{j=1}^{n}(\alpha_j X_j + \beta_j)\right] = \sum_{j=1}^{n}\alpha_j \mathbb{E}X_j + \sum_{j=1}^{n}\beta_j$$

- Variance is *not linear* and *in-variate to shifting*.

$$\mathbb{V}\mathrm{ar}\left[\sum_{j=1}^{n}(\alpha_j X_j + \beta_j)\right] = \sum_{j=1}^{n}\alpha_j^2\ \mathbb{V}\mathrm{ar}(X_j)$$

In words, expectation distributes across sums and constants factor out of expectation. Variation is unaffected by shifting and constants factor out squared. These properties are easily derived from the definitions, hence left to the reader as an exercise.

example 1.7 Toss a fair coin twice. Let $X =$ number of heads in the two tosses. Compute $\mathbb{E}X$ and $\mathbb{V}\mathrm{ar}(X)$.

solution 1.7 Recall $f_X(x) = \frac{1}{4}\delta(x-0) + \frac{1}{2}\delta(x-1) + \frac{1}{4}\delta(x-2)$. Hence,

$$\begin{aligned}
\mathbb{E}X &= \int_{\mathbb{R}} x f_X(x) dx \\
&= \int_{\mathbb{R}} x\left[\frac{1}{4}\delta(x-0) + \frac{1}{2}\delta(x-1) + \frac{1}{4}\delta(x-2)\right] dx \\
&= \frac{1}{4}(0) + \frac{1}{2}(1) + \frac{1}{4}(2) \\
&= 1
\end{aligned}$$

Similarly for variance,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(X) &= \int_{\mathbb{R}} (x-\mu)^2 f_X(x) dx \\
&= \int_{\mathbb{R}} (x-\mu)^2\left[\frac{1}{4}\delta(x-0) + \frac{1}{2}\delta(x-1) + \frac{1}{4}\delta(x-2)\right] dx
\end{aligned}$$

$$= \frac{1}{4}(0-1)^2 + \frac{1}{2}(1-1)^2 + \frac{1}{4}(2-1)^2$$

$$= \frac{1}{2}$$

One could've just as easily apply the definition given earlier for $\mathbb{E}g(X)$, instead of utilizing $\delta(\,\cdot\,)$. Perhaps it would be more straight-forward for $\mathbb{E}X$, whereas it is more expedient to use $\delta(\,\cdot\,)$ for $\mathbb{V}\mathrm{ar}(X)$.

1.9 Common Probability Distributions

This section has been pulled from the STAT-134 Brief Treatises on Discrete Probability and Continuous Probability, John-Michael Laurel. It's meant to be a quick overview of named probability distributions and their properties.

1.9.1 Discrete Uniform Distribution

If $U \sim$ **Uniform** $\big(\{a, a+1, \ldots, b\}\big)$

$$\mathbb{P}(U = u) = \frac{1}{n}$$

That is, the chance of getting any u is the same. $\mathbb{E}U = \frac{a+b}{2}$, $\mathbb{V}\mathrm{ar}(U) = \frac{(b-a+1)^2-1}{12}$ e.g. rolling a fair n-sided

1.9.2 Bernoulli Distribution

If $X_p \sim$ **Bernoulli** (p), then $X_p \in \{0, 1\}$. Here, 1 corresponds to success and 0 failure; with probabilities p and $1-p$ respectively. $\mathbb{E}X_p = p$ and $\mathbb{V}\mathrm{ar}(X_p) = p(1-p)$. e.g. flipping a p coin

1.9.3 Binomial Distribution

A generalization of one Bernoulli trial. i.e., the sum of n IID **Bernoulli** (p) trials. For $\big(X_j(p)\big) \sim$ **Bernoulli** (p), then

$$S_n(p) = \sum_{j=1}^{n} X_j(p) \sim \textbf{Binomial}\,(n, p)$$

and for the event $\{S_n(p) = k\}$ successes

$$\mathbb{P}\big(S_n(p) = k\big) = \binom{n}{k} p^k (1-p)^{n-k}$$

$\mathbb{E}S_n = np$ and $\mathbb{V}\mathrm{ar}(S_n) = np(1-p)$ e.g. getting k heads in n independent p coin flips

1.9.4 Poisson Distribution

A limiting distribution of **Binomial** (n, p) where $np \to \mu$ as $n \uparrow \infty$ and $p \downarrow 0$. In words, we have many trials and the event of success is rare. Via consecutive probability ratios and for $N_\mu \sim$ **Poisson** (μ), we derive

$$\mathbb{P}(N_\mu = k) = \mathbb{P}(0) \prod_{i=1}^{k} R(i) = e^{-\mu}\, \frac{\mu^k}{k!}$$

where $R(i) := \frac{\mathbb{P}(i)}{\mathbb{P}(i-1)}$. $\mathbb{E}N_\mu = \mathbb{V}\mathrm{ar}(N_\mu) = \mu$. e.g. number of Twitter notifications

1.9.5 Geometric Distribution

Another extension of **Bernoulli** (p), where we yield success on the k^{th} trial, implying exactly $k-1$ failures before that. For $G_p \sim$ **Geometric** (p) supported on $\{1, 2, \ldots\}$, we have

$$\mathbb{P}(G_p = k) = (1-p)^{k-1} p$$

One could also compute the probability of success happening after k trials, which is logically

equivalent to not succeeding in the first k,

$$\mathbb{P}(G_p > k) = (1-p)^k$$

We interpret the Geometric distribution as describing the number of trials up to and including the first success. $\mathbb{E}G_p = \frac{1}{p}$ and $\mathbb{V}\mathrm{ar}(G_p) = \frac{1-p}{p^2}$ e.g. tossing a p coin until you get a head

In the situation where $G'_p \sim$ **Geometric** (p) is supported on $\{0, 1, \dots\}$, we have

$$\mathbb{P}(G'_p = k) = (1-p)^k p$$

and

$$\mathbb{P}(G'_p > k) = (1-p)^{k+1}$$

In this context, we are counting the number of failures before yielding a success. $\mathbb{E}G'_p = \frac{1-p}{p}$ and $\mathbb{V}\mathrm{ar}(G'_p) = \frac{1-p}{p^2}$.

Elaboration on support. We think of G_p and G'_p as *shifted* versions of one another. To illustrate this notion, suppose we toss a p-coin, then for $G_p \sim$ **Geometric** (p), we can realize the event $\{G_p = 4\}$ as a sequence of zeroes and ones.

$$(0, 0, 0, 1)$$

Observe for $G'_p \sim$ **Geometric** (p), $\{G'_p = 3\}$ has this same realization

$$(0, 0, 0, 1)$$

These follow immediately from the PMFs of G_p and G'_p. Observe the event $\{G'_p = k-1\}$ translates to "$k-1$ failures before the first success." That is, we toss tails $k-1$ times and on the k^{th} toss, we succeed in tossing heads. The event $\{G_p = k\}$ translates to "yielding success on the k^{th} trial." That is, we toss tails for the first $k-1$ tosses and on the k^{th} toss, we succeed in tossing heads. With this in mind, it's obvious that

$$G_p = G'_p + 1$$

The "shifted" description should make sense now. If it doesn't, read the section again and think harder.[2] If it does make sense, awesome! It's super obvious why $\mathbb{V}\mathrm{ar}(G_p) = \mathbb{V}\mathrm{ar}(G'_p)$.

[2] Perhaps with another person.

1.9.6 Negative Binomial Distribution

A generalization of **Geometric** (p), where we wait for the r^{th} success. Naturally, if the r^{th} success happens on the k^{th} trial, this implies that in the first $k-1$ trials we've had exactly $r-1$ successes. For $G_r(p) \sim$ **NegativeBinomial** (r, p), where T_r denotes the number of trials until the r^{th} success, we have

$$\mathbb{P}\big(G_r(p) = k\big) = \binom{k-1}{r-1} p^{r-1}(1-p)^{k-r} \times p$$

$\mathbb{E}(T_r) = \frac{r}{p}$ and $\mathbb{V}\mathrm{ar}(X) = \frac{r(1-p)}{p^2}$ e.g. tossing a p coin until you get the r^{th} head

1.9.7 Hypergeometric Distribution

The analog to **Binomial** (n, p) where trials are dependent. Suppose you have a population of size N with G good elements and B bad elements. You collect a sample of $n \leqslant N$ elements without replacement and wish to know the chance of getting g good elements. For $S_n \sim$ **Hypergeometric** (n, N, G),

$$\mathbb{P}(S_n = g) = \frac{\binom{G}{g}\binom{B}{b}}{\binom{N}{n}}$$

$\mathbb{E}S_n = n\left(\frac{G}{N}\right)$, $\mathbb{V}\mathrm{ar}(S_n) = n\left(\frac{G}{N}\right)\left(\frac{B}{N}\right)\left(\frac{N-n}{n-1}\right)$ e.g. chance of getting 3 aces in a hand of 13 cards dealt from a standard deck

1.9.8 Multinomial Distribution

A generalization of **Binomial** (n, p), where instead of two categorical outcomes[3] (success or failure), we have k. Let $X_j = x_j$ to be the number of occasions of category j, where $N := x_1 + x_2 + \cdots + x_n$ and $p_j := \mathbb{P}(X_j = x_j)$. We

[3] synonyms: classes, bins

then have the following prescription for the joint distribution

$$\mathbb{P}\left(\bigcap_{j=1}^{n}\{X_j = x_j\}\right) = \binom{N}{x_1, \ldots, x_n}\prod_{j=1}^{n} p_j^{x_j}$$

Just as in Binomial, the probabilities (p_j), sum to unity. i.e. $\sum_{j=1}^{n} p_j = 1$. To communicate that the *random vector* $(X_1, \ldots, X_k)$ is distributed Multinomial, we write

$$(X_1, \ldots, X_n) \sim \textbf{Multinomial}\,(N, \vec{p})$$

Where $\vec{p}$ is a *probability vector*[4]. It's easily shown that the marginal distribution of each $X_j \sim \textbf{Binomial}\,(N, p_j)$, reinforcing our intuition. e.g. finding the probability of getting 1 A, 3 B's, 5 C's, 15 D's, and 10 F's from a class, where

Letter Grade Count

grade	A	B	C	D	F
count	15	22	10	32	21

1.9.9 Continuous Uniform Distribution

A uniform random variable X on the interval (a, b), denoted $X \sim \textbf{Uniform}\,(a)\,b$ has density

$$f_X(x) = \frac{1}{b-a}$$

1.9.10 Exponential Distribution

The continuous analog to $G_p \sim \textbf{Geometric}\,(p)$. If $\mathcal{E}_\lambda \sim \textbf{Exponential}\,(\lambda)$, then

$$f_{\mathcal{E}_\lambda}(t) = \lambda e^{-\lambda t}$$

for $t \geqslant 0$ and $\lambda > 0$ a *rate*. We interpret $\mathcal{E}_\lambda$ as the waiting time before some arrival. The right tail probability, called the *survival function*

$$\mathbb{P}(\mathcal{E}_\lambda > t) = e^{-\lambda t}$$

gives the probability that $\mathcal{E}_\lambda$ survives beyond t. In this context, $\mathcal{E}_\lambda$ is the waiting time until some end; usually death. $\mathbb{E}\mathcal{E}_\lambda = \frac{1}{\lambda}$, $\mathbb{V}\mathrm{ar}(\mathcal{E}_\lambda) = \frac{1}{\lambda^2}$

Memoryless Property. $\textbf{Exponential}\,(\lambda)$ is characterized by the *memoryless property*. Framing $\mathcal{E}_\lambda \sim \textbf{Exponential}\,(\lambda)$ as the lifetime of some entity, the distribution of the its remaining life after t has the same distribution of when time started. More formally, for $t, \Delta t > 0$

$$\mathbb{P}(\mathcal{E}_\lambda > t + \Delta t \mid \mathcal{E}_\lambda > t) = \mathbb{P}(\mathcal{E}_\lambda > \Delta t)$$

You can think of the remaining lifetime of a light-bulb as being exponentially distributed. It's chance of dying has the same distribution throughout its lifetime. In other words, it's utility is just as good as it was when it first turned on.

1.9.11 Gamma Distribution

A generalization of $\mathcal{E}_\lambda$. Suppose we define $\{W_i\}_{i=1}^{r} \sim \textbf{Exponential}\,(\lambda)$ independent, then $T_r = \sum_{i=1}^{r} W_i \sim \textbf{Gamma}\,(r, \lambda)$ and

$$f_{T_r}(t) = \frac{\lambda^r}{(r-1)!}\, t^{r-1} e^{-\lambda t}$$

Showing the above requires a quick re-brief on $\textbf{Poisson}\,(\mu)$. Define $N_t \sim \textbf{Poisson}\,(\lambda t)$ as the number of arrivals in time t with rate $\lambda > 0$. It then follows,

$$\mathbb{P}(N_t = k) = e^{-\lambda t}\,\frac{(\lambda t)^k}{k!}$$

We now have enough to derive $\mathbb{P}(T_r \in dt)$.

Proof. First observe $\mathbb{P}(T_r \in dt)$ is equivalent to

$$\mathbb{P}(N_t = r-1,\ \text{arrival in } dt)$$

which for $W \sim \textbf{Exponential}\,(\lambda)$ can be expressed as

$$\mathbb{P}(N_t = r-1)\mathbb{P}(W \in dt \mid N_t = r-1)$$

The first factor is cake and the second is simply λdt. Stitching everything together,

$$\mathbb{P}(T_r \in dt) = e^{-\lambda t}\,\frac{(\lambda t)^{r-1}}{(r-1)!} \times \lambda dt$$

□

[4] a *probability vector* is one whose entries sum to unity

1.9.12 Gaussian Distribution

The continuous analog and approximation to **Binomial** (n, p) for n large and p not close to 0 or 1. If $Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then

$$\phi_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \textbf{exp}\left[-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right]$$

The distribution's parameters μ and σ^2 are $\mathbb{E}Z$ and $\mathbb{V}\mathrm{ar}(Z)$ respectively. See digram on the next page.

1.9.13 Standard Normal Distribution

When $\mu = 0$ and $\sigma^2 = 1$, we have the *standard normal distribution*. So if $X \sim \mathcal{N}(0, 1)$, then

$$\phi_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

We often transform a random variable into a standard normal via a linear change of scale. So for Y not in standard units, we standardize Y

$$Y^* = \frac{1}{\sigma}Y - \frac{\mu}{\sigma}$$

, hence $\mathbb{E}Y^* = 0$ and $\mathbb{V}\mathrm{ar}(Y^*) = 1$. In words, we subtract the mean and divide by the standard deviation.

1.9.14 Linear Combination of Normals

Let $\{X_j\}_{j=1}^n \sim \mathcal{N}\left(\mu_j, \sigma_j^2\right)$ independent and for $\alpha_j \in \mathbb{R}$, define $S_n = \sum_{j=1}^n \alpha_j X_j$, then

$$S_n \sim \mathcal{N}\left(\sum_{j=1}^n \alpha_j \mu_j, \sum_{j=1}^n \alpha_j^2 \sigma_j^2\right)$$

That is, the sum of normals is still normal. For the special case of $\{X_j\}_{j=1}^n \sim \mathcal{N}(0, 1)$ independent and $\sum_{j=1}^n \alpha_j^2 = 1$, we know

$$S_n \sim \mathcal{N}(0, 1)$$

by *spherical symmetry*. It can be easily verified that $\mathbb{E}S_n = 0$ and $\mathbb{V}\mathrm{ar}(S_n) = 1$. Parameterizing the constants in a linear combination of $X, Y \sim \mathcal{N}(0, 1)$, namely $X_\theta = \cos(\theta)X + \sin(\theta)Y$ showcases this idea[5]. Now suppose $X, Y \sim \mathcal{N}(0, 1)$ independent, then

$$X^2 + Y^2 \sim \textbf{Exponential}\left(\frac{1}{2}\right)$$

a miracle result.

1.9.15 Rayleigh Distribution

A distribution with no parameters. Use when dealing with circles. Suppose $T \sim$ **Exponential** $\left(\frac{1}{2}\right)$ and $R = \sqrt{T}$, then $R \sim$ **Rayleigh** with density for $r \in \mathbb{R}^+$

$$f_R(r) = re^{-\frac{1}{2}r^2}$$

and CDF

$$F_R(r) = 1 - e^{-\frac{1}{2}r^2}$$

We also conclude for X and Y defined in **§(1.9.14)** that $\sqrt{X^2 + Y^2} \sim$ **Rayleigh**.

[5] since we're in $\mathbb{R}^2$, we then call this *rotational symmetry*

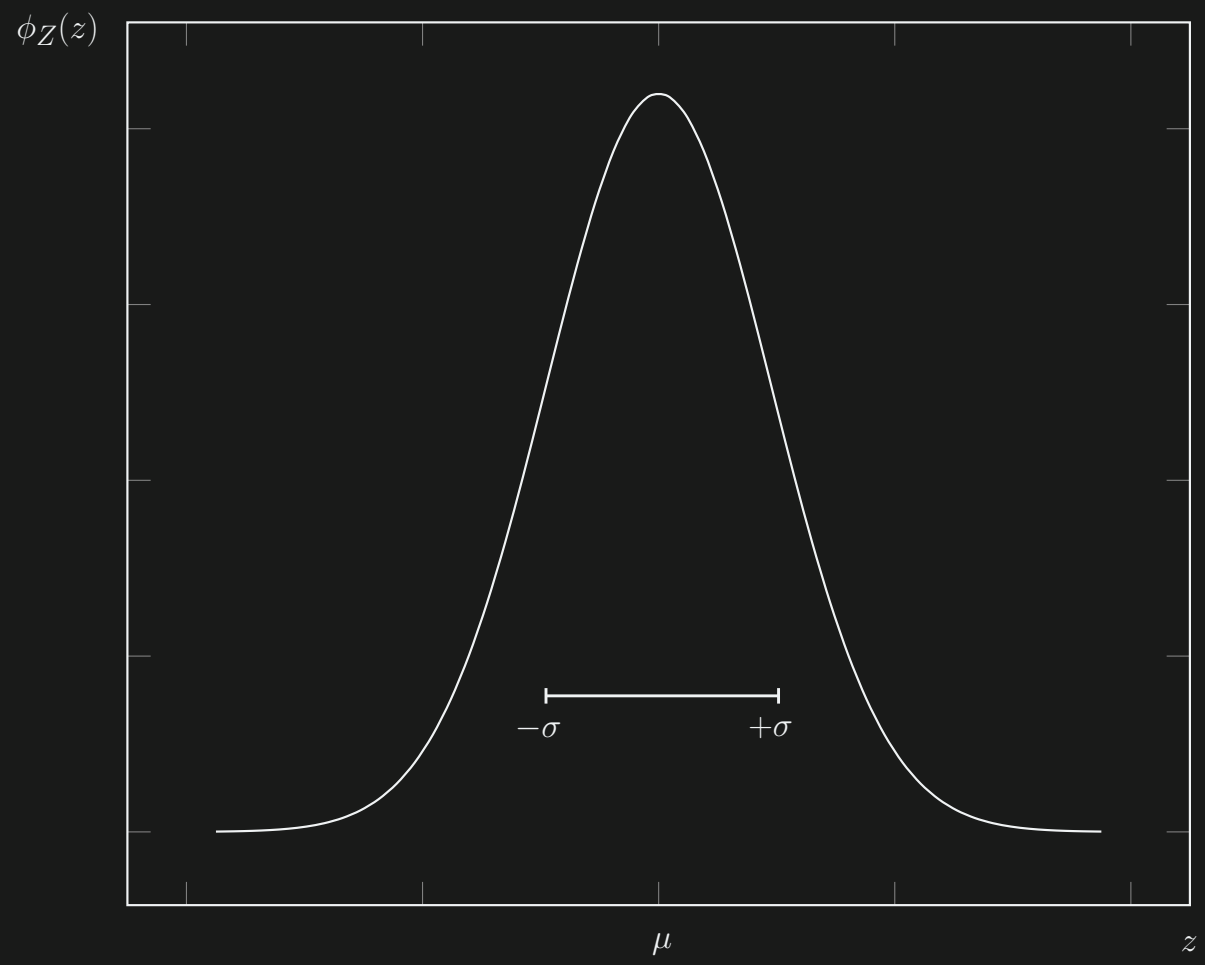

Figure 1.12 Gaussian Distribution with dispersion. This distribution is completely characterized by its mean and variance. The single most important property of Gaussian distributions is that a linear combination of independent Gaussians is itself Gaussian

LECTURE 2

Method of Moments

Imagine a scenario in which we are building an engineering model for a telephone call center to handle airline reservations. For the purpose of analysis, we need to determine (i) the average time spent handling a single call, and (ii) the variance of time spent handling a single call. To determine these quantities, we decide to conduct an experiment in which we record the length of time spent handling n randomly chosen calls over the time span of one month. The question is how can we determine the average and variance using this data?

2.1 Setting the Stage Abstractly

We can abstract this scenario into the following mathematical setting: Suppose $X_1, X_2, \ldots, X_n$ are IID random variables from some unknown distribution with cdf $F_X(x)$. We should think of the X_j as n independent measurements from a single unknown distribution, and this corresponds to the length of time spent handling the randomly chosen calls in our scenario above. The mathematical question we are interested in answering is how to determine $\mu(X)$ and $\mathbb{V}\mathrm{ar}(X)$?

2.2 Law of Large Numbers

One useful insight comes from the *Law of Large Numbers* (LLN). Roughly speaking, the LLN states that if $X_1, X_2, \ldots, X_n$ are IID random variables, then the sample average of the X_j converges to the true mean $\mathbb{E}X$ as n tends towards infinity. Informally stated, For $X_1, X_2, \ldots, X_n$ IID random variables, as $n \uparrow \infty$.

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mathbb{E}X$$

In words, the *sample mean* approaches the *true mean* of X. Stated more precisely,

John-Michael Laurel. Thursday, January 23, 2020

(Weak) Law of Large Numbers

For $X_1, X_2, \ldots, X_n$ IID random variables,

$$\lim_{n\uparrow\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n} X_j - \mu(X)\right| \geqslant \epsilon\right) = 0$$

for $\epsilon > 0$.

All the above is saying is there's a small chance that the sample mean deviates from the true mean of X. This idea sits on the foundation of the frequency interpretation of probability.

2.2.1 Mean and Variance Estimates

From this we can use the *sample average* $\frac{1}{n}\sum_{j=1}^{n} X_j$ as an *estimate* for the true mean $\mathbb{E}X = \mu(X)$. We introduce the following 'hat' notation for the *estimate* a quantity. For example, the sample average is denoted

$$\hat{\mu} := \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Similarly for variance and after some algebra, we derive

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^2,$$

since the variance of X is equal to the expected value of X^2 minus the square of its expected value.

example 2.1 Suppose the data from the call center experiment is $X_1 = 3, X_2 = 1, X_3 = 20, X_4 = 6, X_5 = 5$. In a real experiment, we would obviously have more than $n = 5$ measurements. Compute the mean and variance estimates.

solution 2.1 Then using the above formulas, our estimate of the mean and variance are

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{5} X_j = 7$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^2 = 45.2.$$

2.2.2 Typical Call Center Model

In the above call center scenario, we assumed that the distribution for the time spent handling a single call was unknown. However, it is the case that the distribution for the service time in a call

center is often well represented by an *exponential distribution*. Recall $\mathcal{E}_\lambda \sim$ **Exponential** (λ) with *intensity* $\lambda > 0$ iFF its PDF is

$$f_{\mathcal{E}_\lambda}(t) = \lambda e^{-\lambda t} \mathbb{1}_{\{t \geqslant 0\}}.$$

Integrating, we push out the CDF of $\mathcal{E}_\lambda$

$$F_{\mathcal{E}_\lambda}(t) = (1 - e^{-\lambda t}) \mathbb{1}_{\{t \geqslant 0\}}.$$

After some computation, one finds $\mathbb{E}\mathcal{E}_\lambda = \frac{1}{\lambda}$ and $\mathbb{V}\mathrm{ar}(\mathcal{E}_\lambda) = \frac{1}{\lambda^2}$. This obviously implies $\mathbb{V}\mathrm{ar}(\mathcal{E}_\lambda) = \mathbb{E}^2\mathcal{E}_\lambda$. Coming back to the call center scenario, the estimators for mean and variance that we derived above are not guaranteed to satisfy this relationship. In fact, from the model above, we have:

$$\hat{\mu}^2 = 49 \quad \text{and} \quad \hat{\sigma}^2 = 45.2$$

This is not a good situation, because it means that we are not using the full statistical knowledge available to us when making the estimates of mean and variance. Given this mismatch, we turn our attention towards the question of how we can develop mean and variance estimators that make better use of our statistical knowledge.

2.3 Abstract Model: Estimating the Parameters of a Known Distribution

We can abstract this scenario into the following mathematical setting:

Mathematical Setting for Estimating Parameters of a Known Distribution

Suppose $X_1, X_2, \ldots, X_n$ are IID random variables from some *known* distribution with common CDF

$$F_X(x; \theta_1, \theta_2, \ldots, \theta_p)$$

where $\theta_1, \theta_2, \ldots, \theta_p$ are the *shape paramters* of the distribution.

For concreteness, here are two examples:

- $Z \sim \mathcal{N}(\theta_1, \theta_2)$, where $\mathbb{E}Z = \theta_1$ and $\mathbb{V}\mathrm{ar}(Z) = \theta_1$ and
- $\mathcal{E} \sim$ **Exponential** (θ_1), where $\mathbb{E}\mathcal{E} = \frac{1}{\theta_1}$.

The next natural question is "How best to estimate $\theta_1, \theta_2, \ldots, \theta_p$, given $X_1, X_2, \ldots, X_n$?" In other words, given the n samples, how can one best estimate their parameters? This is not an unreasonable problem, coming back to the call center scenario, if one assumes the call lengths are exponentially distributed, then estimating the rate of the distribution will give the estimate of the mean and variance. So instead of trying to estimate the mean and variance independently for the call center scenario, we shift our perspective to seeing the data coming from an exponential distribution with some unknown rate, estimate that rate, and then estimate the mean and variance. This brings us to the *Method of Moments*. It's perhaps one of the simplest of methods for estimating parameters of a distribution.

2.4 Method of Moments Estimator

Let $\mathbb{E}X^\kappa$ be the κ^{th} *moment for* X. For notational simplicity, define $\mu_\kappa := \mathbb{E}X^\kappa$. The *method of moments estimator* (MoM) is conducted using the following procedure.

Method of Moments

Step 1
Symbolically compute the first p moments, expressing them as functions of the shape parameters $\theta \in \mathbb{R}^p$.

$$\begin{cases} \mu_1 = \mathbb{E}X \ = \mu_1(\theta_1, \theta_2, \ldots, \theta_p) \\ \mu_2 = \mathbb{E}X^2 = \mu_2(\theta_1, \theta_2, \ldots, \theta_p) \\ \quad \vdots \\ \mu_p = \mathbb{E}X^p = \mu_p(\theta_1, \theta_2, \ldots, \theta_p) \end{cases}$$

Step 2
Invert the system of equations. That is, instead of the moments in terms of the parameters, have the parameters in terms of the moments.

$$\begin{cases} \theta_1 = \theta_1(\mu_1, \mu_2, \ldots, \mu_p) \\ \theta_2 = \theta_2(\mu_1, \mu_2, \ldots, \mu_p) \\ \quad \vdots \\ \theta_p = \theta_p(\mu_1, \mu_2, \ldots, \mu_p) \end{cases}$$

Step 3
Motivated by LLN, compute the following estimates of moments

$$\begin{cases} \hat{\mu}_1 = \frac{1}{n}\sum_{j=1}^n X_j \\ \hat{\mu}_2 = \frac{1}{n}\sum_{j=1}^n X_j^2 \\ \quad \vdots \\ \hat{\mu}_p = \frac{1}{n}\sum_{j=1}^n X_j^p \end{cases}$$

Step 4
Compute estimates of the parameters by substituting $\hat{\mu}_\kappa$ for μ_κ in the set of functions $\theta_1, \theta_2, \ldots, \theta_p$ which gives the following estimators of the parameters:

$$\begin{cases} \hat{\theta}_1 = \theta_1(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_p) \\ \hat{\theta}_2 = \theta_2(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_p) \\ \quad \vdots \\ \hat{\theta}_p = \theta_p(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_p) \end{cases}$$

Stepping back from the above procedure, you should be able to see that the estimator $\hat{\sigma}^2$ from the call center example is a precursor of the MoM. In particular, we expressed $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}X^2 - \mathbb{E}^2X$. In words, we expressed variance in terms of the first and second moments of X.

example 2.2 Estimate the parameter for $\mathcal{E}_{\theta_1} \sim$ **Exponential** (θ_1) using MoM.

solution 2.2 Since $\mathcal{E}$ has one parameter $p = 1$. Proceeding with the MoM,

Step 1. Computing the first moment $\mu_1 := \mathbb{E}X = \dfrac{1}{\theta_1}$.

Step 2. Inverting the above, $\theta_1 = \dfrac{1}{\mu_1}$.

Step 3. An estimator for the first moment is the sample average $\hat{\mu}_1 = \frac{1}{n}\sum_{j=1}^{n} X_j$.

Step 4. Substituting $\hat{\mu}_1$ into $\hat{\theta}_1(\mu_1) = \hat{\theta}_1(\hat{\mu}_1) = \dfrac{1}{\frac{1}{n}\sum_{j=1}^{n} X_j}$.

Turning back to the call center scenario, we can push the values $X_1 = 3, X_2 = 1, X_3 = 20, X_4 = 6, X_5 = 5$ into θ_1 and get

$$\hat{\theta}_1 = \hat{\theta}_1(\hat{\mu}_1) = \frac{1}{7}$$

Estimating the mean and variance, we substitute the estimated parameter into the equations for the mean and variance, we get

$$\hat{\mu} = \frac{1}{\hat{\theta}_1} = 7 \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{\hat{\theta}_1} = 49$$

Comparing this estimated mean and variance to the previous estimates, we see that the estimated mean remains the same but the estimated variance is larger. However, this set of estimates is arguably better because these estimates satisfy the relationships we would expect if the data is drawn from an exponential distribution. That is, in concordance to $\mathcal{E}_\lambda \sim$ **Exponential** (λ), we have $\mathbb{E}^2\mathcal{E}_\lambda = \mathbb{V}\mathrm{ar}(\mathcal{E}_\lambda)$.

example 2.3 Suppose $T_r^{(1)}, T_r^{(2)}, \ldots, T_r^{(n)}$ are a supply of IID Gamma distributed random variables with common density

$$f_{T_r}(t) = \frac{1}{\Gamma(r)\theta^r} t^{r-1} e^{-\frac{t}{\theta}}.$$

Compute the MoM estimators for r and θ.

solution 2.3 We'll stick with the given parameters for the density given above (as opposed to mapping $r = \theta_1$ and $\theta = \theta_2$) for notational simplicity.

Step 1. Since T_r is Gamma distributed, we obviously have

$$\mu_1 = r\theta \quad \text{and} \quad \mu_2 = r\theta^2 + (r\theta)^2.$$

Step 2. Computing $\theta_\kappa = \theta_\kappa(\mu_1, \mu_2)$, we get

$$\theta = \frac{\mu_2 - \mu_1^2}{\mu_1} \quad \text{and} \quad k = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

Step 3. For $\kappa = 1, 2$, the estimators are (uninterestingly),

$$\hat{\mu}_\kappa = \frac{1}{n}\sum_{j=1}^{n}\left(T_r^{(j)}\right)^\kappa$$

Step 4. Constructing $\hat{k}(\hat{\mu}_1, \hat{\mu}_2)$ and $\hat{\theta}(\hat{\mu}_1, \hat{\mu}_2)$, we have (in gory detail),

$$\hat{k}(\hat{\mu}_1, \hat{\mu}_2) = \frac{\left(\frac{1}{n}\sum_{j=1}^{n} T_r^{(j)}\right)^2}{\frac{1}{n}\sum_{j=1}^{n}(T_r^{(j)})^2 - \left(\frac{1}{n}\sum_{j=1}^{n} T_r^{(j)}\right)^2}$$

$$\hat{\theta}(\hat{\mu}_1, \hat{\mu}_2) = \frac{\frac{1}{n}\sum_{j=1}^{n}(T_r^{(j)})^2 - \left(\frac{1}{n}\sum_{j=1}^{n} T_r^{(j)}\right)^2}{\frac{1}{n}\sum_{j=1}^{n} T_r^{(j)}}$$

The example above highlights the bottleneck in performing MoM. That is, step 2 tends to require the most friction, since the system of equations generated are non-linear, hence requires more algebra. We can relate this to the method of Lagrangian multipliers from our multivariate calculus course. You set up the system of non-linear equations and hope to see the best way to proceed. That is, there's no real systematic way of solving, you just do it.

2.4.1 Philosophical Discussion

Why is MoM liked by economists and not so much by statisticians? Economists like this method because it's easy, in the sense that for a very complicated model, there's an algorithm that a person can solve and run to get answers to the question. The *Generalized Method of Moments* was developed to further push the MoM. They like it because one can specify very complex models and in principle there's an algorithm for how to get the parameters of a model. Statisticians don't like it MoM, because of step 1 and step 3. We estimate the κth moment by taking the sample average of X_j^κ. Now imagine you had a hundred parameters, imagine what it would be like to take for example X_j^{100}. Numerically it's terrible, but there is a statistical issue as well. As soon as you start taking powers of your data, it is as if you are amplifying the *uncertainty* in the data. The whole issue of statistics is to say something about the real distributions. But the problem is, we don't know what the real distribution is. So the data we start with already represents uncertainty. The problem with the MoM is that in taking the data and raising them to powers, we are actually amplifying the uncertainty in the model. Generally, the MoM does not produce the best estimates for the data collected. So why are we learning it? It's really more to get our feet wet and to familiarize ourselves with the MoM, which is the most fundamental method for parameter estimation. But it won't be the method we employ when performing actual analysis, instead we'll be applying more sophisticated approaches for analysis.

2.5 Linear Model of Building Energy

The amount of energy E_i consumed in a building on the ith day heavily depends on the outside temperature on the ith day T_i. Generally, there are two situations. When the outside temperature is low, decreasing temperature leads to increased energy consumption because the building needs to provider greater amounts of heating. When the outside temperature is high, increasing temperature leads to increased energy consumption because the building needs to provide greater amounts of cooling. This is a general phenomenon, and is empirically observed with the real data shown in Figure 2.1, which was collected from Sutardja Dai Hall.

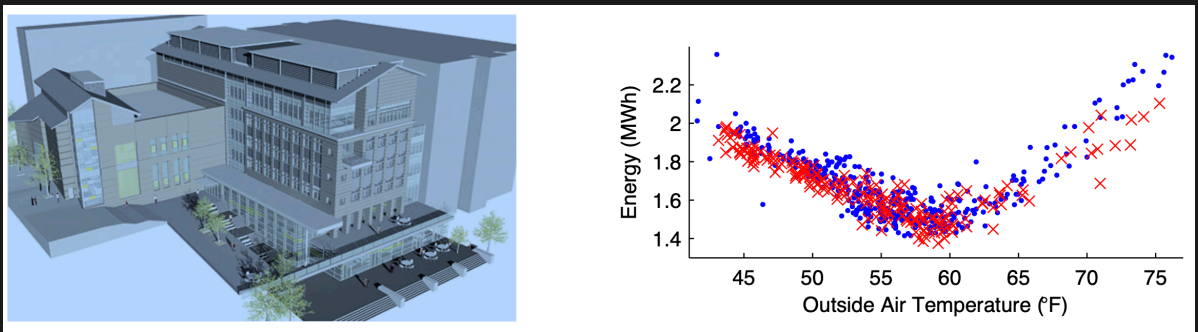

Figure 2.1 Real data (right) from Sutardja Dai Hall (left) is shown. The data marked with dots and crosses denote measurements made running two different versions of software on the heating, ventilation, and air-conditioning (HVAC) system in Sutardja Dai Hall. This type of graph is called a *scatter plot*. This v shape trend is very classical in real data sets.

Based on the figure, we might guess that when the outside temperature is below 59°F, the relationship between energy consumption and outside temperature is linear and given by

$$E = \alpha T + \beta + \epsilon,$$

where $\alpha, \beta \in \mathbb{R}$ unknown and ϵ is a zero mean, finite variance random variable, and is independent of T. The random variable ϵ is appended to capture the "noise" of the model; it is an allowance of uncertainty. That is, no model will entirely capture a situation. So for example, humidity was a parameter not in consideration to the linear model or the number of people throughout the day. Even if you were to capture all the parameters, there's inherent randomness with the data itself. Class example: YouTube videos versus Word documents. Nonetheless, ϵ captures uncertainty. What is the method of moments estimate of $\alpha, \beta, \mathbb{V}\mathrm{ar}(\epsilon)$ if we have n measurements of (T_i, E_i)? There are two useful hints for this problem: (i) Use the random variable $\epsilon = E - \alpha T - \beta$ (since we have a distribution, ostensibly) and (ii) use $\mu_1(\epsilon)$ and $\mu_2(T\epsilon)$ in step 1, instead of $\mu_1(\epsilon)$ and $\mu_2(\epsilon)$. Proceeding with the four steps:

Step 1. Observe $\mu_1(\epsilon) = \mu_1(T\epsilon) = 0$ since $\mathbb{E}\epsilon = 0$ and independent of T, hence the moments are

$$\mu_1(\epsilon) = \mu_1(E) - \alpha\mu_1(T) - \beta = 0$$
$$\mu_1(T\epsilon) = \mu_1(ET) - \alpha\mu_2(T) - \beta\mu_1(T) = 0$$

For clarity on the second line $\mu_2(T) := \mathbb{E}T^2 = \mu_1(T^2)$.

Step 2. The above produces a linear systems of equations,

$$\begin{bmatrix} \mu_1(T) & 1 \\ \mu_2(T) & \mu_1(T) \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \mu_1(E) \\ \mu_1(ET) \end{bmatrix} \implies$$

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \frac{1}{\big(\mu_1(T)\big)^2 - \mu_2(T)} \begin{bmatrix} \mu_1(T) & -1 \\ -\mu_2(T) & \mu_1(T) \end{bmatrix} \begin{bmatrix} \mu_1(E) \\ \mu_1(ET) \end{bmatrix} \implies$$

$$\alpha\big(\mu_1(E), \mu_1(T), \mu_1(ET), \mu_2(T)\big) = \frac{\mu_1(E)\mu_1(T) - \mu_1(ET)}{\big(\mu_1(T)\big)^2 - \mu_2(T)}$$

$$\beta\big(\mu_1(E), \mu_1(T), \mu_1(ET), \mu_2(T)\big) = \frac{\mu_1(ET)\mu_1(T) - \mu_1(E)\mu_2(T)}{\big(\mu_1(T)\big)^2 - \mu_2(T)}$$

Step 3. The estimators of the moments are (uninterestingly),

$$\hat{\mu}_1(E) = \frac{1}{n}\sum_{j=1}^{n} E_j \qquad \hat{\mu}_1(T) = \frac{1}{n}\sum_{j=1}^{n} T_j$$

$$\hat{\mu}_1(ET) = \frac{1}{n}\sum_{j=1}^{n} E_j T_j \qquad \hat{\mu}_2(T) = \frac{1}{n}\sum_{j=1}^{n} T_j^2$$

Step 4. Shove $\hat{\mu}_1(E), \hat{\mu}_1(T), \hat{\mu}_1(ET)$, and $\hat{\mu}_2(T)$ into α and β and you're done, that is you get an expression for $\hat{\alpha}$ and $\hat{\beta}$. Using the sample average notation, i.e., the bar notation, we have

$$\hat{\alpha} = \frac{\overline{E}\,\overline{T} - \overline{ET}}{\overline{T}^2 - \overline{T^2}} \quad \text{and} \quad \hat{\beta} = \frac{\overline{ET}\,\overline{T} - \overline{E}\,\overline{T^2}}{\overline{T}^2 - \overline{T^2}}$$

It's worthy note that the above is a clumsy derivation for the estimators α and β. It's truly tedious, but if you follow the steps of the MoM algorithm, we can get $\hat{\alpha}$ and $\hat{\beta}$ for something that, initially was thought to not be applicable to. In summary, we have a linear model with unknown coefficients and we can use the MoM to estimate those parameters. Future lectures will address a more robust and simpler approach for constructing such *linear* models.

University of California, Berkeley

Department of Industrial Engineering and Operations Research

John-Michael Laurel

Expanded lecture notes adapted from Anil Aswani's spring 2020 lecture compilation for IEOR-165 course on engineering statistics, quality control, and forecasting.

LECTURE NOTES