Contents

1 Fundamentals

Note: material in this section is largely taken from Chapters 1-3 of Wasserman's book ("All of Statistics")

1.1 Sample Spaces, Events, Probabilities

Imagine these two situations:

Evil Ruler. An evil ruler has imprisoned three innocent students (Alice, Bob, and Carol), visiting from a foreign country. The three are in separate cells and unable to communicate. As a show of force, he has decided to execute two of the three, in ten days. As a show of mercy, he has decided that the captives will not know their fates, except to know that two of the three will be executed. The guards have been instructed accordingly.

Alice doesn't like her odds. But she is sly and says to her guard: "I know that Bob, or Carol, or both will be executed. Give me the name of one of them that will be executed. I cannot communicate with them, so you are telling me nothing about my fate and breaking no rules."

To the guard, this seems reasonable and he answers "Bob is one of the two."

Alice's spirits improve, as she calculates that her chances of being executed have improved from $\frac{2}{3}$ to $\frac{1}{2}$. After all, she and Carol are equally likely to be the remaining victim.

Money Machine. I fill an envelope with an amount of money (you don't know how much) and I fill a second envelope with twice the amount in the first.

I choose one of the two envelopes at random and hand it to you. Before you open it, I offer to exchange it for the other envelope.

Being a well-trained and sensible data scientist, you make a statistical calculation:

Let X be the amount in the envelope that you have in hand. As for the other envelope, it has

$$Y = \begin{cases} 2X & \text{with probability } p = \frac{1}{2} \\ \\ \frac{X}{2} & \text{with probability } p = \frac{1}{2} \end{cases}$$

and your expected windfall is $E[Y] = E[X]$ if you keep your envelope, and $E[Y] = \frac{1}{2}E[X] + \frac{1}{2}E[\frac{X}{2}] = 1.25E[X]$ if you make the exchange. It's a no-brainer: 25% expected gain for an exchange that costs nothing. Oddly, the same argument would then apply to the new envelope in hand, which is a bit confusing since it argues for yet another exchange. But that can't be right. After all, you'd end up where you started—with the original envelope.

Without further concerns, you make the exchange.

Something is not right, in both situations. That something is the lack of a coherent statistical model. Let's remember how to build one.

1.1.1 Sample Spaces

Definition. *A sample space, Ω, is the collection of all possible outcomes of an experiment:*

$$\omega \in \Omega \leftrightarrow \textit{one possible outcome}$$

e.g. Flip two coins: $\Omega = \{HH, HT, TH, TT\}$

e.g. Evil ruler:

$$\begin{aligned}\Omega &= \{AB, AC, BC\} \times \{\text{guard says } B, \text{guard says } C\} \\ &= \{ABb, ABc, ACb, ACc, BCb, BCc\}\end{aligned}$$

where, for example, ACc is shorthand for the event "Alice and Carol will be killed, and the guard said 'Carol'," and ABc is shorthand for the event "Alice and Bob will be killed, and the guard said 'Carol"' (which is notably unlikely, unless we assume that the guard lies, but that's an issue for later in the modeling—when we assign probabilities).

e.g. Money machine: The trick is to come up with a sample space that "covers the bases," meaning that it includes all of the information that we would need to determine the relevant quantities. In this case, the relevant quantities are the amounts in the envelope that I first gave to you, and the amount in the one that I've offered in exchange.

Let's call the first envelope (the one to which I first add a sum of money) "Envelope 1," and the other envelope (the one to which I add twice the amount) "Envelope 2." I will use a to represent the amount I add to Envelope 1 and 's' to represent whether I first hand you Envelope 1 (say, $s = 0$) or Envelope 2 (say, $s = 1$). Then (a, s) will clearly be sufficient for determining the quantities of interest. In other words,

$$\Omega = \big\{(a, s) : a \in (0, \infty), s \in \{0, 1\}\big\} \tag{1}$$

is a suitable sample space.

1.1.2 Events

Definition. *Events are subsets of the sample space. Sometimes, the set of events needs to be restricted. In general we will use $\mathcal{E}$ to represent the collection of (allowed) events.*

When the possible outcomes are represented by a discrete set (i.e. Ω is at most countably infinite[1], then we will always take $\mathcal{E} = 2^\Omega$, meaning the set of all subsets of Ω.

e.g. Flip two coins: $\mathcal{E} = 2^\Omega =$

$$\begin{aligned}\big\{&\varnothing, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \\ &\{HT, TH\}, \{HT, TT\}, \{TH, TT\}, \{HH, HT, TH\}, \{HH, HT, TT\}, \\ &\{HH, TH, TT\}, \{HT, TH, TT\}, \{HH, HT, TH, TT\}\big\}\end{aligned}$$

[1] can be counted—put in one-to-one correspondence with the positive integers

e.g. Evil ruler: $\mathcal{E} = 2^{\Omega}$, which consists of 32 subsets of Ω (why?)

e.g. Money machine: I suggested using $\Omega = (0, \infty) \times \{0, 1\}$ (this is just another, more compact, way to write equation (1)). This time, it turns out that 2^{Ω} is just too big, and we have to content ourselves with something considerably more modest. The reason is subtle. I will have a little more to say about this shortly. But as it turns out, the consequence is irrelevant (unless you're contemplating a drastic change in career direction following your masters degree).

1.1.3 Probabilities

Definition. *Given $(\Omega, \mathcal{E})$, a probability is a function $\mathbb{P} : \mathcal{E} \to [0, 1]$ with the following three properties:*

i. $\mathbb{P}(A) \geq 0 \quad \forall\, A \in \mathcal{E}$

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

iii. If $A_1, A, \ldots$ are disjoint events in $\mathcal{E}$ ($A_i \cap A_j = \varnothing,\ \forall\, i \neq j$) then

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Remarks

1. $(\Omega, \mathcal{E}, \mathbb{P})$ is called a probability model.
2. N.B.[2] Models are almost always (if not in fact always) an approximation or idealization, and never unique. That's why they're called models.
3. If $A, B, A_1, A_2, \ldots, A_n \in \mathcal{E}$, then we will often write $\mathbb{P}(AB)$ as shorthand for $\mathbb{P}(A \cap B)$ and $\mathbb{P}(A_1 A_1 \cdots A_n)$ as shorthand for $\mathbb{P}(\cap_{i=1}^{n} A_i)$.

e.g. Flip two coins: $\mathbb{P}(\omega) = \frac{1}{4}\ \forall \omega \in \Omega$

So, $\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}) = \frac{1}{4}$. Then, by applying property (iii), we get $\mathbb{P}$ on all of $\mathcal{E}$.

e.g. $\mathbb{P}(\{HH, TT\}) = \mathbb{P}(\{HH\}) + \mathbb{P}(\{TT\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

e.g. $\mathbb{P}(\text{"first flip is tails"}) = \mathbb{P}(\{TH, TT\}) = \mathbb{P}(\{TH\}) + \mathbb{P}(\{TT\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

e.g. Evil ruler: perhaps the most natural model is to assume that (i) each of the three possible pairs of executions (i.e. AB, AC, and BC) are equally likely, (ii) the guard doesn't lie, and (iii) when faced with a choice (in the 'BC' case) of saying "B will die" or "C will die" the guard chooses at random, 50/50. This leads to the following model:

[2] Nota Bene, note well

$$\begin{array}{lccccccc} \Omega = \{ & \text{ABb} & \text{ABc} & \text{ACb} & \text{ACc} & \text{BCb} & \text{BCc} & \} \\ \mathbb{P} = & \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & \end{array}$$

Now that we have a model, we're almost ready to reexamine Alice's reasoning, and perhaps resolve any discomfort over her apparent ability to lower her odds of execution by receiving information of no apparent value. Indeed, once we've made precise the notion of a conditional probability (as in, say, "the conditional probability that Alice will be executed given that her guard answered 'Bob"') the resolution will be immediate. Stay tuned.

e.g. Troubles in the continuum: Let $\Omega = [0, 1]$. There is no reason, *a priori*, to restrict $\mathcal{E}$, so let's assume, for the time being, that $\mathcal{E} = 2^\Omega$, the set of all subsets of the unit interval. A good probability to experiment with might be the *uniform* probability, which we might reasonably expect to have the following two properties (among others):

(*) ("generalization of length") $\mathbb{P}(A) = b - a$ when $A = [a, b]$, for any $0 \leq a \leq b \leq 1$; and

(**) ("shift invariance") For any $A \subseteq [0, 1]$ and $x \in [0, 1]$, $\mathbb{P}(A) = \mathbb{P}\big((A + x) \mod 1\big)$, where $(A + x) \mod 1$ just shifts A by x and then wraps the portion bigger or equal to one around to zero, as though zero were connected to one to make a circle out of the interval $[0, 1]$. For example, we would agree that under the uniform probability we want $\mathbb{P}([0.5, 1]) = \mathbb{P}([0, 0.25] \cup [0.75, 1])$ (where I've shifted $[0.5, 1]$ by $x = 0.25$).

Now here's something not obvious: it turns out that there is *no function* $\mathbb{P} : 2^\Omega \to [0, 1]$ that simultaneously satisfies (*), (**), and the properties (i), (ii), (iii) . The set of subsets of $[0, 1]$ is just too big. At least no such function unless we are willing to modify the usual axioms from which mathematical theory is derived. We know how to do this, but there is a heavy price to pay, so we leave it alone.

But the news is far from all bad. We can satisfy all five conditions ((*), (**), (i), (ii), (iii)) on a very large and entirely adequate collection of subsets (see, for example, the "Borel sets," or the larger collection called the "Lebesgue sets"). We will leave the construction of $\mathcal{E}$ to a course in measure theory, but with the reassurance that $\mathcal{E}$ is big enough to include anything that will come up in practice.

Point: we can't always take $\mathcal{E}$, the set of events, to be the set of *all* subsets of Ω, but can assume that $\mathcal{E}$ has all of the events of interest in any practical application.

The Rolling Stones got it right: "You can't always get what you want, but if you try sometimes you find you get what you need."

e.g. Money machine: Here are two (of many) possible probability models. We can use the same sample space, namely the one set up in equation (1), for each model.

Model 1 For (almost[3]) any $A \subseteq (0, \infty)$

$$\begin{aligned}\mathbb{P}(a \in A, s = 0) &= \mathbb{P}(a \in A, s = 1) \\ &= \frac{1}{2} \int_{x \in A} e^{-x} dx\end{aligned}$$

We will get to the strict definition of *independence* soon, but for now think of the model as describing a process of (1) choosing a value of a to put into Envelope 1, and then (2) independently choosing which of the two envelopes (Envelope 1 with amount a, or Envelope 2 with amount $2a$) to hand to you, according to the outcome of a fair coin flip (probability $\frac{1}{2}$ for each of the two possibilities).

Model 2 This is a discrete model, which assumes that I put an integer amount of money in Envelope 1, namely k dollars with probability $\frac{1}{2^k}$, for any $k = 1, 2, \ldots$. The choice of which envelope to hand to you is again independent of the amount. Thus, in a formulation equivalent to that of Model 1: For any $A \subseteq (0, \infty)$

$$\begin{aligned}\mathbb{P}(a \in A, s = 0) &= \mathbb{P}(a \in A, s = 1) \\ &= \frac{1}{2} \sum_{\{k \in \mathbb{Z}^+ \colon k \in A\}} \frac{1}{2^k}\end{aligned}$$

where $\mathbb{Z}^+$ denotes the positive integers.

The basic three properties of a probability have many ramifications. Here are just a few. The ones that are not derived (namely 2 and 3) make for good exercises.

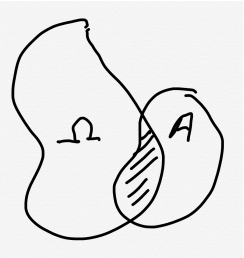
1. $\mathbb{P}(\varnothing) = 0$ $\big($because: $1 = \mathbb{P}(\Omega) = \mathbb{P}(\Omega \cup \varnothing) = \mathbb{P}(\Omega) + \mathbb{P}(\varnothing) = 1 + \mathbb{P}(\varnothing)\big)$
2. $A, B \in \mathcal{E} \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
3. Given $A_1, A_2, \ldots \in \mathcal{E}$, let $B = \cup_{k=1}^{\infty} A_k$, and $C = \cap_{k=1}^{\infty} A_k$. Then
   (*) $\lim_{n \to \infty} \mathbb{P}(\cup_{k=1}^{n} A_k) = \mathbb{P}(B)$ ("continuity from below")
   (*) $\lim_{n \to \infty} \mathbb{P}(\cap_{k=1}^{n} A_k) = \mathbb{P}(C)$ ("continuity from above")
4. If $\Omega = [0, \infty)$ and $A_i = [i, \infty)$ $\forall i = 1, 2, \ldots$, then for *any* probability $\mathbb{P}$, $\lim_{i \to \infty} \mathbb{P}(A_i) = 0$ $\big($because: $\cap_{i=1}^{\infty} A_i = \varnothing$ and $A_i = \cap_{k=1}^{i} A_k$, so, by continuity from above, $\mathbb{P}(A_i) = \mathbb{P}(\cap_{k=1}^{i} A_k) \to \mathbb{P}(\varnothing) = 0\big)$

We end this section by looking at two more examples of constructing uniform distributions on Ω:

[3] It's silly to keep saying 'almost every' whenever we're talking about subsets of the continuum, so let's agree that, starting now, we will simply say "for all $A \subseteq (0, \infty)$" or "for all $A \subseteq (0, 1)$," or for any other continuum (uncountable) set, knowing full well that it isn't quite true.

e.g. The uniform probability on a set in the plane: Here, $\Omega \subseteq \mathbb{R}^2$ is a region in the plane, and for any subset $A \subseteq \mathbb{R}^2$ we define

$$\mathbb{P}(A) \triangleq \frac{\text{Area}(A \bigcap \Omega)}{\text{Area}(\Omega)}$$



e.g. The uniform distribution on committees: Ω is all subsets of size 3 out of 100 objects (e.g. all committees of three people that can be formed out of 100 people). If $|\Omega|$ represents the size of Ω (e.g. the number of distinct committees), then

$$|\Omega| = \binom{100}{3} = \frac{100!}{97!3!} = \frac{100 \cdot 99 \cdot 98}{6} = 1{,}717$$

Since Ω is countable (in fact, finite), we can take $\mathcal{E} = 2^{\Omega}$, and for any $\omega \in \Omega$ define $\mathbb{P}(\omega) = \frac{1}{1{,}717}$, which extends to all of $\mathcal{E}$ (property (iii) of probabilities) to give us the *uniform distribution* on Ω.

1.1.4 Independence and Conditional Probabilities

Definition. *The sets $A, B \in \mathcal{E}$ are independent (denoted $A \perp\!\!\!\perp B$) if*

$$\mathbb{P}(AB) = \mathbb{P}(A)\,\mathbb{P}(B)$$

(recall that $\mathbb{P}(AB)$ is shorthand for $\mathbb{P}(A \bigcap B)$)

Remark By definition, $\varnothing$ and Ω are each independent of every set in $\mathcal{E}$ (why?).

e.g. Money machine: Independence of the amount I placed in Envelope 1 (which we called a) from the choice of the envelope that I handed to you (which we represented by $s \in \{0, 1\}$) was built in. In the first model, for example, for any $A \subseteq (0, \infty)$ and $s = 0$,

$$\mathbb{P}(a \in A, s = 0) = \frac{1}{2} \int_A e^{-x} dx = \mathbb{P}(a \in A)\,\mathbb{P}(s = 0)$$

e.g. If $A \perp\!\!\!\perp B$, $B \perp\!\!\!\perp C$, and $A \perp\!\!\!\perp C$ ("pairwise independence"), then is

$$\mathbb{P}(ABC) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)?$$

(See homework, where you will be asked to either prove this or construct a counter example.)

Definition. *If $\mathbb{P}(B) > 0$ then the conditional probability of A given B is denoted by $\mathbb{P}(A|B)$ and defined by*

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

Remark Conditional probabilities are probabilities! Given $B \in \mathcal{E}$ with positive probability, define

$$\mathbb{P}_B(A) = \mathbb{P}(A|B) \quad \forall A \in \mathcal{E}$$

Now convince yourself that $\mathbb{P}_B$ is a (new) probability on $\mathcal{E}$, i.e. it satisfies the three properties of a probability. Conditioning is just a matter of defining a new, smaller, "universe"—B replaces Ω.

This is obvious, but handy, too. For example, from the definition of conditional probability, it is clear that

$$\mathbb{P}(AB) = \mathbb{P}(A|B)\,\mathbb{P}(B) \quad \forall\, A, B \text{ with } \mathbb{P}(B) > 0 \tag{2}$$

If we focus now on the "universe" of B, then for any C with $\mathbb{P}(C) > 0$ we conclude that

$$\mathbb{P}(AC|B) = \mathbb{P}(A|CB)\,\mathbb{P}(C|B)$$

by the same rule as (2), but applied to what we called $\mathbb{P}_B$, above. Wrap your mind around this one—it is sure to save you time.

We're ready to end any remaining mystery (if there is any left to end) about Alice's all-too-easy reduction in her odds of being executed: "the conditional probability that Alice will be executed given that her guard answered 'Bob"' is just $\mathbb{P}(\text{Alice dies}|\text{guard says 'B'})$, which we can read off directly from our model:

$$\frac{\mathbb{P}(ABb)}{\mathbb{P}(b)} = \frac{\mathbb{P}(ABb)}{\mathbb{P}(ABb) + \mathbb{P}(BCb)} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{6}} = \frac{2}{3}$$

As for the more subtle "Money Machine," we still need more machinery, which will be developed in the following two sections on Random Variables and Expectations.

But before moving on, let's do another exercise in conditional probabilities. This one is typical of a calculation that is commonly used in the interpretation of medical tests. It is about the so-called PSA ("prostate-specific antigen") test for prostate cancer in men. The test measures the concentration of PSA in urine, and generally uses a cutoff of 4 ng/ml (nanograms per milliliter): higher concentrations are considered a positive result, possibly indicating prostate cancer. (Hardly a typical use of the word "positive.")

Let $+$ indicate a positive test result and $-$ a negative test result (concentration at or below 4 ng/ml). Furthermore, let $\mathcal{C}$ indicate the presence of prostate cancer (as verified, say, with a biopsy) and let $\sim \mathcal{C}$ indicate the absence of prostate cancer.

According to the recent literature, $\mathbb{P}(+|\mathcal{C}) \approx 0.21$ (this is called the "sensitivity" of the test), and $\mathbb{P}(-|\sim \mathcal{C}) \approx 0.91$ (called the "specificity" of the test).

Now suppose that a patient tests positive. Naturally, he will be interested in knowing the probability that he has prostate cancer (at least he *should* be interested). This is a *conditional* probability, conditioning on his positive test result. How do we go about calculating $\mathbb{P}(\mathcal{C}|+)$? Reason as follows:

$$
\begin{aligned}
\mathbb{P}(\mathcal{C}|+) &= \frac{\mathbb{P}(+,\mathcal{C})}{\mathbb{P}(+)} \\
&= \frac{\mathbb{P}(+|\mathcal{C})\,\mathbb{P}(\mathcal{C})}{\mathbb{P}(+,\mathcal{C}) + \mathbb{P}(+,\sim\mathcal{C})} \\
&= \frac{\mathbb{P}(+|\mathcal{C})\,\mathbb{P}(\mathcal{C})}{\mathbb{P}(+|\mathcal{C})\,\mathbb{P}(\mathcal{C}) + \mathbb{P}(+|\sim\mathcal{C})\,\mathbb{P}(\sim\mathcal{C})} \\
&= \frac{\mathbb{P}(+|\mathcal{C})\,\mathbb{P}(\mathcal{C})}{\mathbb{P}(+|\mathcal{C})\,\mathbb{P}(\mathcal{C}) + \big(1-\mathbb{P}(-|\sim\mathcal{C})\big)\big(1-\mathbb{P}(\mathcal{C})\big)} \\
&= \frac{0.21\,\mathbb{P}(\mathcal{C})}{0.21\,\mathbb{P}(\mathcal{C}) + \big(1-0.91\big)\big(1-\mathbb{P}(\mathcal{C})\big)} \\
&= \frac{0.21\,\mathbb{P}(\mathcal{C})}{0.09 + 0.12\,\mathbb{P}(\mathcal{C})}
\end{aligned}
$$

The point is that the calculation of $\mathbb{P}(\mathcal{C}|+)$ has been reduced to a simple formula that depends only on the *incidence* of prostate cancer in the population being tested, i.e. it depends only on $\mathbb{P}(\mathcal{C})$. In fact, as you can plainly see, the approach is generic: given the sensitivity and specificity of a test, there is a formula for the probability of having the associated disease that depends only on the incidence of the disease in the population being tested.

Suppose that our patient is a twenty-two year-old male. It turns out that approximately 5% of the male population below the age of thirty do have prostate cancer. (This is an alarmingly high number, but most prostate cancers are very slow growing and, in fact, will never be diagnosed and are unlikely to be the cause of death. The 5% estimate is based on autopsies of men under thirty, who will typically have died from other causes). For our patient, then, we would take the incidence to be, approximately, $\mathbb{P}(\mathcal{C}) = 0.05$. According to our formula, the patient's probability of having prostate cancer, given his positive PSA result, is about 0.11.

On the other hand, men who are 79 years old or older have an incidence of approximately 0.59 ($\mathbb{P}(\mathcal{C}) \approx 0.59$), in which case a positive test is more concerning, with $\mathbb{P}(\mathcal{C}|+) \approx 0.77$ (derived, again, by just plugging into the formula).

Here's a more dramatic example: A recent test developed for rapid screening for HIV positive individuals has sensitivity 0.775 and specificity 0.993 ($\mathbb{P}(+|\mathcal{H}) = 0.775,\ \mathbb{P}(-|\sim\mathcal{H}) = 0.993$, where I have used $\mathcal{H}$ to indicate HIV infection. Mimicking the above calculations, but for the HIV rapid screening test, gives the following expression for the probability of being HIV positive given a positive result on the test:

$$
\mathbb{P}(\mathcal{H}|+) = \frac{0.775\,\mathbb{P}(\mathcal{H})}{0.007 + 0.768\,\mathbb{P}(\mathcal{H})}
$$

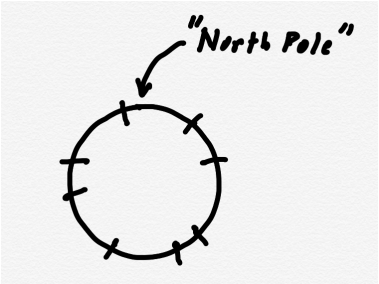
In a high-risk population (say IV street-drug users), the incidence could be as high as $\mathbb{P}(\mathcal{H}) = 0.1$, in which case $Prob(\mathcal{H}|+) \approx 0.92$. But in a lower-risk general population the incidence could be as low as one in a thousand, $\mathbb{P}(\mathcal{H}) = 0.001$, or even lower. In this case, despite the very low false-positive rate of

the rapid screening test ($\mathbb{P}(+| \sim \mathcal{H})$=1-0.993=0.007), $\mathbb{P}(\mathcal{H}|+) < 0.1$. An individual from this population who tests positive has a chance of less than one in ten of actually carrying the virus.

There is a simple "back-of-the-envelope" calculation that makes a good approximation and can be quite instructive. Take this last example and imagine that we test a sample of 1,000 individuals from the low-risk pool. Typically we will have about one infected individual from this group (since the incidence is only 0.001), and about 1,000 uninfected. Let's assume that the test turns up positive for the infected individual (the probability, in any case, is 0.775). About how many other positive tests should we expect? Of the *uninfected* individuals, the test will produce about (0.007)·(1000)=7 positive results. This reasoning suggests that only about one out of eight positive tests come from infected individuals, which isn't so far off from the more precise calculation that gave us one out of ten.

1.2 Random Variables

e.g. Suppose we place eight points randomly (uniformly) and independently on a circle with circumference equal to 1, as in the figure. This creates eight intervals, all of which have the same distribution on (arc) length (soon, we will call this "identically distributed"). As we shall see, the "expected" length of each is $\frac{1}{8}$, which should not be too much of a surprise. But this might surprise you: If we consider the length of the interval that contains the "North Pole," **its** expected length is **bigger** (in fact $\frac{2}{9}$—which is not what I said in class, but does have the advantage of being correct).



This has to do with *conditional* expectations, and the fact that conditioning on being *the* interval containing the North Pole (or any other directions you might prefer, so long as it chosen before you look at the intervals) is evidence for being larger than the others!

More generally we lay down n points, the expected length of a randomly chosen interval is $\frac{1}{n}$, but the particular one that contains the north pole has expected length $\frac{2}{n+1}$. Not only that, but the other $n-1$ intervals got shorter: their expected lengths are now $\frac{1}{n+1}$. Kind of odd. Let's develop some machinery so that we can make sense of it.

e.g. Every one arrives late to a mixer. A proper name tag was prepared in advance for each of the n people, but in the rush the tags are handed out randomly. Assuming that the names are unique, what is the expected number of people who get the right name tag?

1.2.1 Definition and Examples

Definition. *A random variable, X, is a function $X : \Omega \to \mathbb{R}$.*

(This is a bit of a lie. Not *all* functions are allowed, at least not when Ω is not countable. But it's the same story (ultimately for the same reason), and even the conclusion is the same: The issue will *never* come up in practice.)

e.g. Flip a coin four times. $\Omega = \{H,T\}^4 = \{\omega_1,\omega_2,\omega_3,\omega_4 : \omega_i \in \{H,T\} \;\; i = 1,2,3,4\}$. If our only interest is in the number of heads, then we can focus the particular function $X(\{\omega_1,\omega_2,\omega_3,\omega_4) = $ # heads.

e.g. Roll a die twice.

$$\begin{aligned}\Omega &= \big\{(n,m) : n \in \{1,2,\ldots,6\}, m \in \{1,2,\ldots,6\}\big\} \\ &= \{1,2,\ldots,6\}^2 \\ &= \{1,2,\ldots,6\}^{\{1,2\}} \\ &= \{1,2,\ldots,6\} \times \{1,2,\ldots,6\}\end{aligned}$$

(to name just a few of the ways to write the same thing!). If our interest is in the *sum* of the two outcomes, as it often is when playing with dice, then we might define the random variable $X(n,m)$ to be $n+m$.

e.g. Money machine: $\big\{(a,s) : a \in (0,\infty), s \in \{0,1\}\big\}$, where $s = 0$ indicates that I give you Envelope 1, and $s = 1$ indicates that I give you Envelope 2, and a is the amount placed in Envelope 1. There are several random variables of interest, e.g. $A(a,s) = a$, the amount in Envelope 1, and $S(a,s) = s$, the indicator of which envelope you have been handed. Probably, your greatest interest is in $X(a,s)$, the amount in the envelope that you've been handed: $X(a,s) = (1-s)\cdot a + s \cdot 2a = a(1+s)$.

In general, functions of random variables are themselves random variables. For example, we could just as well have written X, in the money machine example, in terms of the random variables A and S, as $X = A\cdot(1+S) = A(a,s)\cdot(1+S(a,s))$. And there is yet another frunction on Ω of particular interest: the amount in the *other* envelope, the one not in your hand: $Y = S\cdot A + (1-S)\cdot 2A = A\cdot(2-S)$.

1.2.2 Distribution Functions, Mass Functions and Density Functions

Mostly, our interest is in random variables and not in Ω, *per se*. Examples include the *fraction* of people polled who will vote for candidate D, and not (say) the number of those people who are male or female, or old or young; the amount of money you will make with a particular portfolio of stocks, and not necessarily the individual stocks that make up the portfolio; the number of heads in a sequence of coin tosses, and necessarily the sequence of heads and tails.

The probabilities of possible values of a random variable are fully captured by its *cumulative distribution function* (cdf):

Definition. *The cdf of a random variable X is the function*

$$F_X(x) = \mathbb{P}(X \le x) \;\; \forall x \in \mathbb{R}$$

or, more explicitly,

$$F_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \le x\})$$

Remarks

1. It is a good discipline, and common practice, to use an upper case letter for a random variables and a lower case letter for an actual value of the random variable. In other words, an upper case letter for the function and a lower case letter for its value.

2.

$$lim_{x\to-\infty}F(x) = 0 \text{ (by an application of "continuity from above")}$$
$$lim_{x\to\infty}F(x) = 1 \text{ (by an application of "continuity from below")}$$

3. (monotonicity) $x < y \to F(x) \leq F(y)$

4. (F_X contains all probabilistic information about X) If $X \sim F_X$ (meaning X "has the distribution given by" F_X), and $Y \sim F_Y$, and if

$$F_X(z) = F_Y(z) \ \ \forall z \in \mathbb{R}$$

then

$$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A) \quad \text{for all (allowed) } A$$

In this case, we say that "X and Y are identically distributed," and write $X \sim Y$ (X has the same distribution as Y).

e.g. Flip a biased coin twice. The probability of H (heads) on any one flip is 0.2. The natural model has $\Omega = \{HH, HT, TH, TT\}$, four elements, with probabilities $0.2^2 = 0.04$, $0.2 \cdot 0.8 = 0.16$, $0.8 \cdot 0.2 = 0.16$, and $0.8^2 = 0.64$, respectively. Let $X = X(\omega) =$ the number of heads (0, 1, or 2). The cdf is then

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 0.64 & \text{for } 0 \leq x < 1 \\ 0.64 + 0.16 + 0.16 = 0.96 & \text{for } 1 \leq x < 2 \\ 1 & \text{for } x \geq 2 \end{cases}$$

Notice that

(i) $F_X(x)$ is continuous from the right (which can be written as $F_X(x^+) = F_X(x)$, in terms of $F_X(x^+) \triangleq lim_{z\downarrow x}F_X(z)$).

(ii) The sizes of any jumps in F_X at x is the probability $\mathbb{P}(X = x)$ (which can be written as $\mathbb{P}(X = x) = F_X(x) - F_X(x^-)$, in terms of $F_X(x^-) \triangleq lim_{z\uparrow x}F_X(z)$).

e.g. $\Omega = (0, \infty)$, and for any $A \subseteq (0, \infty)$, $\mathbb{P}(A) = \int_{\omega\in A} e^{-\omega}d\omega$. If $X(\omega) = \omega$ then

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{\omega=0}^{x} e^{-\omega}d\omega = (1 - e^{-x})\mathbb{1}_{x>0}$$

e.g. ("change of variables") Same X as in the previous example, but let $Y = 1 + 10X$. Y is a function of X and, as such, is itself a random variable (specifically, in this case, $Y(\omega) = 1 + 10X(\omega) = 1 + 10\omega$).

What is the cdf of Y?

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \le y) \\ &= \mathbb{P}(1 + 10X \le y) \\ &= \mathbb{P}(X \le \frac{y-1}{10}) \\ &= F_X(\frac{y-1}{10}) \\ &= (1 - e^{-\frac{y-1}{10}})\mathbb{1}_{\frac{y-1}{10}>0} \\ &= (1 - e^{-\frac{y-1}{10}})\mathbb{1}_{y>1} \end{aligned}$$

No need to revisit Ω. Everything we needed was in $F_X(x)$!

Definition. *A random variable X is called* discrete *if F_X is flat except at its discontinuities.*

e.g. Flip a coin repeatedly until the first head appears. Assume that the probability of heads on a given flip is 0.4. Let X be the number of flips. The possible outcomes are $\Omega = \{H, TH, TTH, TTTH, \ldots\}$ with individual probabilities $\mathbb{P}(\{H\}) = 0.4$, $\mathbb{P}(\{TH\}) = (0.6)(0.4)$, $\mathbb{P}(\{TTH\}) = (0.6)^2(0.4)$, and, in general, the probability of k tails followed by one head is $(0.6)^k(0.4)$, for $k = 0, 1, 2, \ldots$. Hence

$$F_X(x) = \begin{cases} 0 & \text{for } x < 1 \\ 0.4 & \text{for } 1 \le x < 2 \\ (0.4) + (0.6)(0.4) & \text{for } 2 \le x < 3 \\ (0.4) + (0.6)(0.4) + (0.6)^2(0.4) & \text{for } 3 \le x < 4 \\ \vdots & \end{cases}$$

Or, since $\sum_{k=0}^{n-1} r^k = \frac{1-r^n}{1-r}$ for all $r \in (0, 1)$, $F_X(x) = \frac{1-(0.6)^n}{1-0.6}(0.4) = 1 - (0.6)^n$ whenever $x \in [n, n+1)$, for $n \ge 1$.

A convenient way to capture the probability distribution of a discrete random variable is through its *probability mass function*:

Definition. *The probability mass function (pmf) of a discrete random variable X is the function*

$$f_X(x) \triangleq \mathbb{P}(X = x)$$

Remarks.

1. More explicitly, $f_X(x) = \mathbb{P}(\{\omega : X(\omega) = x\})$, but the idea is to get away from Ω and just work with $f_X(x)$ and $F_X(x)$.
2. As already noted, $\mathbb{P}(X = x_o)$ is equal to the size of the jump in F_X at x_o:

$$f_X(x_o) = F_X(x_o) - \lim_{x \uparrow x_o} F_X(x)$$

Or, for any x, let $x^- = \lim_{y \uparrow x} F_X(y)$. Then $f_X(x) = F_X(x) - F_X(x^-)$.

e.g. Let X be the number of heads in two flips of a biased coin with probability 0.2 of coming up heads:

$$f_X(x) = \begin{cases} 0.64 & \text{for } x = 0 \\ 0.32 & \text{for } x = 1 \\ 0.04 & \text{for } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

e.g. Let X be the number of coin flips until the first appearance of H, for a biased coin with probability 0.4 of coming up H:

$$f_X(x) = \begin{cases} (0.6)^x(0.4) & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Definition. *X is called a* continuous *random variable if there exists a function $f_X : \mathbb{R} \to [0, \infty)$ such that*

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy \quad \forall x \in \mathbb{R}$$

In this case, the function $f_X(x)$ is called the probability density function (pdf).

Remarks.

1. So $F_X(x) = \frac{d}{dx}F_X(x)$ (fundamental theorem of calculus).
2. If X is continuous then
$$\mathbb{P}\left(X \in (a, b)\right) = \int_{x \in (a,b)} f_X(x)dx$$
3. In fact, if X is continuous then
$$\mathbb{P}\left(X \in A\right) = \int_{x \in A} f_X(x)dx$$
4. If X is continuous then
$$\begin{aligned} \mathbb{P}(X = x_o) &\leq \mathbb{P}\left(X \in (x_o - \epsilon, x_o + \epsilon)\right) \\ &= \int_{x \in (x_o - \epsilon, x_o + \epsilon)} f_X(x)dx \\ &\to 0 \text{ as } \epsilon \to 0 \end{aligned}$$
In other words, $\mathbb{P}(X = x_o) = 0$.
5. As a consequence of the observations above:
$$\mathbb{P}\left(X \in (a, b)\right) = \mathbb{P}\left(X \in [a, b)\right) = \mathbb{P}\left(X \in (a, b]\right) = \mathbb{P}\left(X \in [a, b]\right) = \int_{x=a}^{b} f_X(x)dx$$

6. Sometimes $\frac{d}{dx}F_X(x)$ doesn't exist at all x...but it doesn't matter. For example, suppose $X \sim$ uniform on $[0,1]$:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1 \end{cases}$$

Then $\frac{d}{dx}F_X(x)$ is undefined at $x = 0$ and $x = 1$. Still, it is perfectly proper to write $f_X(x) = \mathbb{1}_{x\in[0,1]}$ to represent the pdf of X

7. There are some very strange (maybe exotic is a better word) examples in which $F_X(x)$ is *continuous* for all x but there is no density. For example, there are functions $F_X(x)$ for which: (i) $F_X(0) = 0$, (ii) $F_X(1) = 1$, (iii) $F_X(x)$ is continuous and nondecreasing on $(0,1)$, and (iv) $\frac{d}{dx}F_X(x)$ exists and is zero at every $x \in (0,1)$ except for $x \in D$, where D is so small that $\mathbb{P}(D) = 0$. Yet somehow $F_X(x)$ manages to get from zero to one!

e.g. $\Omega = [0,\infty)$ and for any $A \subseteq [0,\infty)$, $\mathbb{P}(A) = \int_{x\in A} e^{-x}dx$. Suppose that $X(\omega) = \omega$. Then for any $x \geq 0$,

$$\mathbb{P}(X \leq x) = \int_{y=0}^{x} e^{-y}dx = (1 - e^{-x})$$

If $x < 0$ then $\mathbb{P}(X \leq x) = 0$ so we can write, succinctly, $F_X(x) = (1 - e^{-x})\mathbb{1}_{x\geq 0}$. What's more, X is continuous with pdf $\frac{d}{dx}F_X(x) = e^{-x}\mathbb{1}_{x\geq 0}$.

e.g. $F_X(x) = \sqrt{x}\mathbb{1}_{x\in[0,1]}$ (instead of $x\mathbb{1}_{x\in[0,1]}$, as in the $U[0,1]$ distribution). Then

$$f_X(x) = \frac{1}{2}x^{-\frac{1}{2}}\mathbb{1}_{x\in(0,1]}$$

Notice that I've set $f_X(x)$ to zero at $x = 0$. But we're free to use any value there, as is always the case for the density of a continuous random variable at any particular isolated x. Notice also that $f_X(x) \to \infty$ as $x \downarrow 0$, but this does not take away from the fact that f_X is a valid density.

A final note before moving on to random vectors: some distributions are neither discrete nor continuous:

e.g. (a "mixed distribution") $\Omega = [0,\infty)$, and for $A \subseteq [0,\infty)$

$$\mathbb{P}(A) = \frac{2}{3}\int_{\omega\in A} e^{-\omega}d\omega + \frac{1}{3}\mathbb{1}_{2\in A}$$

If $X(\omega) = \omega$, then X is neither a discrete random variable nor a continuous one. But, as is always the case, X does have a cumulative distribution function. See if you can convince yourself that

$$F_X(x) = \frac{2}{3}\left(1 - e^{-x}\right)\mathbb{1}_{x\in[0,\infty)} + \frac{1}{3}\mathbb{1}_{x\geq 2}$$

1.2.3 Multivariate and Marginal Distributions

Definition. *A random vector $X_{1:n} = (X_1, X_2, \ldots, X_n)$ is a collection of random variables all defined on the same probability space $(\Omega, \mathcal{E}, \mathbb{P})$.*

In short, a sequence of random variables.
e.g. The sample space consists of outcomes of ten flips of a coin:

$$\Omega = \{H, T\}^{10} = \{(\omega_1, \ldots, \omega_{10}) : \omega_i \in \{H, T\}, i = 1, \ldots, 10\}$$

and $S_{1:10} = (S_1, \ldots, S_{10})$ where, for each $k = 1, \ldots, 10$, S_k is the number of heads in the first k flips, i.e.

$$S_k = \sum_{i=1}^{k} \mathbb{1}_{\omega_i = H}$$

e.g. The sample space is the ten-dimensional unit cube,

$$\Omega = [0, 1]^{10} = \{(\omega_1, \ldots, \omega_{10}) : \omega_i \in [0, 1], i = 1, \ldots, 10\}$$

and $S_{1:10} = (S_1, \ldots, S_{10})$ where, for each $k = 1, \ldots, 10$

$$S_k = \sum_{i=1}^{k} \omega_i$$

As with single random variables, the distribution of a random vector is fully specified by its cumulative distribution function:

Definition. *The* joint *(aka* multivariate*) cdf of a random vector* $X_{1:n} = (X_1, \ldots, X_n)$ *is*

$$\begin{aligned} F_{X_{1:n}}(x_{1:n}) &= F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \\ &= \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) \\ &= \mathbb{P}(\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\}) \end{aligned}$$

Remark: $F_{X_{1:n}}$ is sometimes called, simply, a *multivariate distribution*.

e.g. $\Omega = [0, 1]^n$, and for every $A \subseteq \mathbb{R}^n$, $\mathbb{P}(A) = \text{Vol}(A \cap [0, 1]^n)$. This is the "uniform probability distribution on the unit cube in $\mathbb{R}^n$."

Given $(\omega_1, \ldots, \omega_n) \in \Omega$, define $X_k(\omega_1, \ldots, \omega_n) = \omega_k$, for each $k = 1, \ldots, n$. So X_k is just the k'th coordinate of a point in the n-dimensional unit cube. As we will see shortly, $X_1, \ldots, X_n$ are then "independent and identically distributed," or iid, random variables. The "identically distributed" part is obvious:

$$\begin{aligned} \mathbb{P}(X_1 \leq x_1) &= \text{Vol}(\{(w_1, \ldots, w_n\} : \omega_1 \leq x_1\}) = x_1 \\ \mathbb{P}(X_k \leq x_1) &= \text{Vol}(\{(w_1, \ldots, w_n\} : \omega_k \leq x_1\}) = x_1 \end{aligned}$$

e.g. Functions of random vectors are random variables. (In fact, functions of random vectors can also be random vectors.) Let $\Omega = [0, 1]^2$, and let $X(\omega_1, \omega_2) = \omega_1$ and let $Y(\omega_1, \omega_2) = \omega_2$. Define a new random variable, from X and Y, as the sum of the two: $Z = X + Y$. Show that the cdf of Z is

$$F_Z(z) = \begin{cases} \frac{z^2}{2} & 0 \leq z \leq 1 \\ 1 - \frac{(2-z)^2}{2} & 1 \leq z \leq 2 \end{cases}$$

(Hint: draw a picture.) Then we can get the pdf by simply differentiating:

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \begin{cases} z & 0 \le z \le 1 \\ 2 - z & 1 \le z \le 2 \end{cases}$$

1. Define joint pmf, pdf and note relationship between pdf and partial derivatives of cdf, give some example F's and f's in 2 dimensions
2. "Going Backwards" - "marginalization" and "independence"
   a. example of getting joint pmf of two discrete random variables from joint pmf of four discrete RV's.
   b. same thing with integrals
3. Independent RV's - motivate by going back to relationship between conditional probabilities and independence - then define independence in terms of factorization of cdf (and equivalence to factorization of pdf or pmf when they exist)
4. Define iid

1.3 Expectation

1.3.1 Definition

for univariate and multivariate rv's, assuming discrete or continuous (i.e. using pmf, pdf)

1.3.2 Linearity

define and use it to solve some otherwise hard problems (like the "name-tag problem" - each of n people get a name tag, but at random. What is expected number that get correct name tag. Use the "indicator trick" and it's easy. Do it "directly" and it's pretty darn hard. (In fact, end up with a nontrivial combinatorial identity by doing it both ways–which I did not derive)

1.3.3 Special Expectations

1. μ_X
2. σ_X^2 and the handy formula $\sigma_X^2 = E[X^2] - \mu_X^2$
3. standard deviation
4. multivariate mean (a vector) $X = (X_1, \ldots, X_n)$, $\vec{\mu}_X = (\mu_{X_1}, \ldots, \mu_{X_n})$
5. covariance
6. covariance matrix, written as
$$C = E[(X - \vec{\mu}_X)^T (X - \vec{\mu}_X)]$$

2 Introduction to Monte Carlo Sampling for Estimation & Hypothesis Testing

Note: much of the material in this section can be found, in greater detail, in Chapter 2 and Chapter 3, Section 3.1 of the draft book "Modern Applications of Probability and Statistics"

2.1 Generating (Pseudo) Random Numbers

Goal: Generate a sequence of independent & identically distributed (iid) random variables. How?

① Use a physical process (decaying isotopes, cosmic rays, quantum mechanics, ...)

② Use a pseudo-random number generator (PRNG) - a deterministic sequence that behaves like an iid sequence of random variables.

[In the past 15-20 years, theory & state-of-the-art for PRNG's have gotten very complex & sophisticated. We will not cover state-of-the-art methods, but we will see what was best 15-20 years ago, and in this way cover the principles & pitfalls underlying all PRNG's.]

Approach:

① Generate a computationally simple deterministic sequence of integers

② Only observe an aspect of the ~~integers~~ sequence that "looks random"

There are many kinds of random variables. Which r.v.'s should we generate? As we will see later, every sequence of r.v.'s can

be generated from an iid sequence 2
of $U(0,1)$ r.v.'s. [$U(0,1) \leftrightarrow X$ with density $f_X(x) = \mathbb{1}_{x \in (0,1)}$, i.e. uniform on the interval $(0,1)$]

If we choose m (integer) very large then we can approximate a $U(0,1)$ r.v. by a discrete r.v. X with

x	$\frac{1}{m}$	$\frac{2}{m}$	.....	$\frac{m-1}{m}$
$p(x)$	$\frac{1}{m-1}$	$\frac{1}{m-1}$		$\frac{1}{m-1}$

i.e. $X \sim$ DRV/$\frac{1}{m}$ $U\{\frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m}\}$
which is the same as using
$\frac{1}{m} X$ when $X \sim U\{1, 2, \ldots, m-1\}$

So our goal is to generate a sequence of r.v.'s that are (approximately) iid $U\{1, 2, \ldots, m-1\}$

2.1.2 Multiplicative Congruential Generators

For any positive integer m, define $x \bmod m = x - \lfloor \frac{x}{m} \rfloor m$ ← round down { remainder after dividing by m

or, equivalently,

$x \bmod m = r \in [0, m)$

where $x = km + r$ for some integer k

Remark: you might want to try to prove the handy formula
(ab)mod m = [(a mod m)(b mod m)] mod m

Remark: Sometimes, $x \bmod m$ refers to the set of all $r+n$, $n=0,1,\ldots$ where r is defined as above.

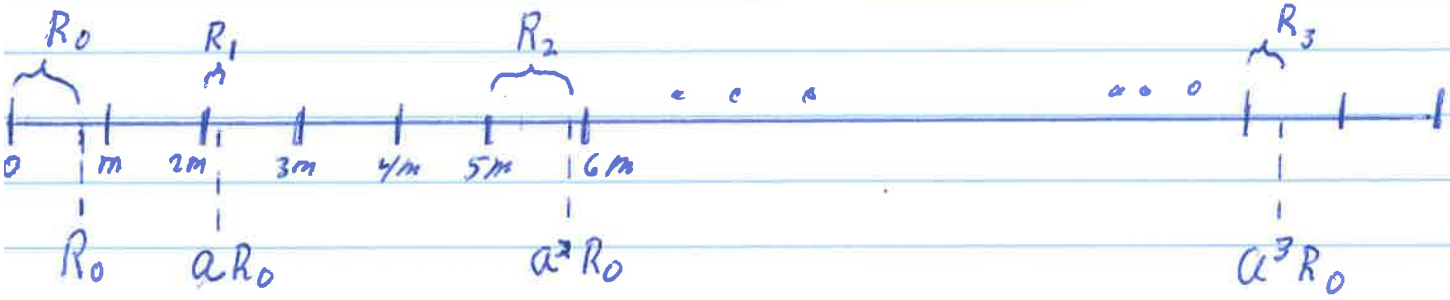
The multiplicative cong. generator is based on

m large prime number

$R_0 \in \{1,2,\ldots,m-1\}$ "seed"

$a \in \{2,3,\ldots,m-1\}$ multiplier (carefully chosen, as explained shortly)

Picture:



$R_0, R_1, R_2, \ldots$ is our sequence of (pseudo) iid $U\{1,2,\ldots,m-1\}$ numbers

Formally: $\boxed{R_n \triangleq (a^n R_0) \bmod m} = (a(a^{n-1}R_0)) \bmod m$

$= ((a \bmod m)((a^{n-1}R_0) \bmod m)) \bmod m = (aR_{n-1}) \bmod m$

↑ the "handy formula"

$a \bmod m = a$

R_{n-1}

Then $R_n \in \{1,2,\ldots,m-1\}$ (proven later) and we use

$\boxed{U_n = R_n/m}$ to approximate $U(0,1)$

e.g. $m=7, R_0=5, a=3$

R	$3R$	$(3R) \bmod 7$
5	15	1
1	3	3
3	9	2
2	6	6
6	18	4
4	12	5
5		
⋮		

So see each of $\{1, \ldots, m-1\}$
"Full Period"

e.g. $m=7, R_0=5, a=2$

R	$2R$	$(2R) \bmod 7$
5	10	3
3	6	6
6	12	5
5		
⋮		

So see only 3 numbers
(not all of $\{1, \ldots, m-1\}$)

not "Full Period" - not
as good

Remarks

* Trick is to choose m and a so that R_n acts like random sequence
* Certainly want full period (no "holes" in (0,1))
* Matlab, before version 5 (currently on version 8.2):

before 1996 { $m = 2^{31}-1$ (prime)
$a = 7^5 = 16,807$ (full period)

But cycle with only $\approx 2^{31} \approx 2$ billion #'s is no good in modern app's that use many more (pseudo) random #'s

* Current Matlab default: "Mersenne Twister"
uses $\{1/2^{53}, 2/2^{53}, \ldots, (2^{53}-1)/2^{53}\}$ with period of $2^{19,936} \approx$ forever on any computer

Theory

1. $R_n \in \{1, 2, \ldots, m-1\}$ $\forall m$ (so $R_n \neq 0$, which would be bad!)

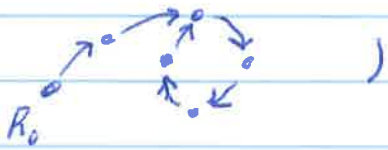
Pf Suppose $R_{n-1} \neq 0$ and $R_n = 0$. Then

$(a R_{n-1}) \bmod m = 0 \Rightarrow a R_{n-1} = km$, some $k > 0$

Factor both sides

$$\underbrace{a R_{n-1}}_{\text{prime factors do not include } m \text{ (since } a \leq m-1 \text{ \& } R_{n-1} \leq m-1)} = \underbrace{km}_{\text{prime factors include } m}$$

2. $R_{m-1} = R_0$ (point: sequence is cyclic, can't have



)

Pf Based on Fermat's "Little Theorem": if p is prime & $q \in \{1, 2, \ldots, p-1\}$ then

$$q^{p-1} \bmod p = 1$$

$$R_{m-1} = (a^{m-1} R_0) \bmod m = [(a^{m-1} \bmod m)(R_0 \bmod m)] \bmod m$$

$$\overset{\uparrow}{=} (R_0 \bmod m) \bmod m \overset{\uparrow}{=} R_0$$

Fermat's — $R_0 \in \{1, 2, \ldots, m-1\}$

Remarks: (1) So full period $\Leftrightarrow$ $R_{m-1} = 1^{st}$ time $R_n = R_0$

3. $R_0 = 1$ has full period $\Leftrightarrow$ $R_0 \neq 1$ has full period, so enough to check by starting at 1.

2.1.2 Evaluating Pseudo-Random Number Generators

Let $U_1, U_2, \ldots$ be a sequence from a PRNG, so $U_k \in (0,1)$ $\forall k$.

We evaluate the PRNG by pretending that

$$U_1, U_2, \ldots \sim \text{iid } U(0,1).$$

What would we expect of a sequence of iid $U(0,1)$ random variables?

Are the expectations met?

With the Law of Large Numbers (LLN)

LLN Suppose $X_1, X_2, \ldots \sim$ iid with common mean μ & variance $\sigma^2 < \infty$

& consider the "sample mean"

$$\bar{X} \triangleq \frac{1}{n}\sum_{k=1}^{n} X_k$$

Then

$$E\bar{X} = \frac{1}{n}\sum_{k=1}^{n} EX_k = \frac{1}{n}\sum_{k=1}^{n} \mu = \mu$$

and

$$V\bar{X} = V\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2} V\left(\sum_{k=1}^{n} X_k\right)$$

$$= \frac{1}{n^2}\sum_{k=1}^{n} V(X_k) = \sigma^2/n \to 0 \quad \text{as } n \to \infty$$

So sample mean is concentrating around μ as $n \to \infty$

(In fact, $P(\lim_{n\to\infty} \bar{X} = \mu) = 1$.)

So for large n, $\bar{X}$ is close to μ with high probability. This is the "law of large numbers," aka the "law of averages."

Back to testing PRNG's. If $U_1, U_2, \ldots$ iid $U(0,1)$, then expect

$$\frac{1}{n}\sum_{k=1}^{n} U_k \to \int_{-\infty}^{\infty} u \, \mathbb{1}_{u\in(0,1)} \, du = \int_0^1 u \, du = .5$$

More generally, expect that for any h

$$h: R^1 \to R^1,$$

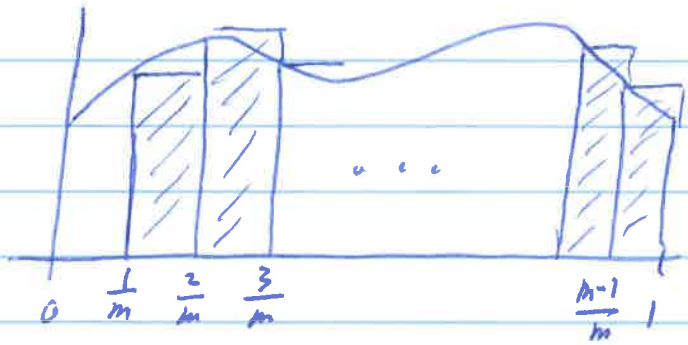
$$\frac{1}{n}\sum_{k=1}^{n} h(U_k) \to E[h(U)] = \int_0^1 h(u)\,du$$

But these are weak tests. For example, consider any multiplicative congruential generator, say based on mod m, some large prime m, and suppose that it has full cycle. When $n = m-1$:

$$\frac{1}{m}\sum_{k=1}^{n} h(U_k) = \frac{1}{m-1}\sum_{k=1}^{m-1} h(U_k) = \frac{1}{m-1}\sum_{k=1}^{m-1} h\left(\frac{k}{m}\right) \underset{m \text{ large}}{\approx} \frac{1}{m}\sum_{k=1}^{m-1} h\left(\frac{k}{m}\right)$$

Riemann sum $\approx \int_{1/m}^{1} h(u)\,du \underset{m \text{ large}}{\approx} \int_0^1 h(u)\,du$

picture:



and when n is very large, say $n = \ell(m-1) + r$ for large ℓ and $r < m-1$, get the same thing

$$\frac{1}{n}\sum_{k=1}^{n} h(U_k) \approx \int_0^1 h(u)\,du$$

$$\left[\frac{1}{n}\sum_{k=1}^{n} h(U_k) \approx \frac{1}{\ell(m-1)}\sum_{k=1}^{n} h(U_k) \approx \frac{1}{\ell(m-1)}\sum_{k=1}^{\ell(m-1)} h(U_k) \right.$$

$$\left. \approx \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{1}{m-1}\sum_{k=1}^{m-1} h(U_k) \approx \frac{1}{\ell}\sum_{i=1}^{\ell} \int_0^1 h(u)\,du = \int_0^1 h(u)\,du \right]$$

$\uparrow$ $U_1, U_2, \ldots$ repeats every $m-1$ samples

So what about stronger tests? How do we actually test the sequence and not just the full-cycle property?

Break the sequence into subsequences, say of length d each, and test the sequence of subsequences:

$(U_1, U_2, \ldots, U_d), (U_{d+1}, U_{d+2}, \ldots, U_{2d}), \ldots,$

$(U_{(n-1)d+1}, U_{(n-1)d+2}, \ldots, U_{nd})$ (n d-dimensional vectors)

These n d-dimensional vectors should look like n iid random vectors, each with Uniform distribution Uniform on

$$(0,1)^d = (0,1) \times (0,1) \times \cdots \times (0,1)$$

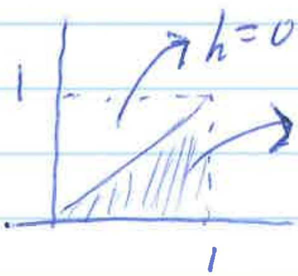
Hence, by the LLN, for any function h of d variables ($h = h(x_1, x_2, \ldots, x_d)$) expect

$$\frac{1}{n} \sum_{k=1}^{n} h(U_{(k-1)d+1}, \ldots, U_{kd}) \to E\, h(U_1, \ldots, U_d)$$

$$= \int_{u_1=0}^{1} \cdots \int_{u_d=0}^{1} h(u_1, u_2, \ldots, u_d)\, du_1\, du_2 \cdots du_d$$

Let's try this with the mult. Cong. Generator using $m = 11$, $a = 2$, $R_0 = 3$, and

$n = 5$, $d = 2$ with $h(u_1, u_2) = \mathbb{1}_{u_1 > u_2}$



$$E\, h(U_1, U_2) = \int_0^1 \int_0^1 h(u_1, u_2)\, du_1\, du_2$$

$$= .5$$

$n = 5$, $d = 2 \Rightarrow 10$ values, $R_0, R_1, \ldots, R_9$ to

$$U_1, U_2, \ldots, U_{10}:$$

	R	$2R$	$(2R) \bmod 11$
R_0	3	6	6
$U_1 \cdot 11 =$	6	12	1
$U_2 \cdot 11 =$	1	2	2
$U_3 \cdot 11 =$	2	4	4
$U_4 \cdot 11 =$	4	8	8
$U_5 \cdot 11 =$	8	16	5
$U_6 \cdot 11 =$	5	10	10
$U_7 \cdot 11 =$	10	20	9
$U_8 \cdot 11 =$	9	18	7
$U_9 \cdot 11 =$	7	14	3
$U_{10} \cdot 11 =$	3		

$$\frac{1}{5}\sum_{k=1}^{5} h(U_{2(k-1)+1}, U_{2k})$$

$= (1+0+1+1+1)/5$

$= 0.8$

Which is not close to the expected value, .5.

(Not looking very random.)

Here's Matlab code (sequence of commands) that does this for $n = 10^6$:

```
>> % parameters of MCG
>> m=11;
>> a=2;
>> R0=3;
>>
>> % parameters of simulation
>> n=10^6;
>> d=2;
>> U=zeros(d,n);
>>
>> % create the pseudo-random sequence
>> R=R0;
>> for k=1:n
     for l=1:d
        R=mod(a*R,m);
        U(l,k)=R/m;
     end
  end
>>
>> % translate to a new random sequence (the 'h' function)
>> H=U(1,:)>U(2,:);
>>
>> % compute the sample mean
>> mean(H)

ans =

   0.8000
```

With hypothesis testing

Let's make our tests more rigorous and precise, using *hypothesis testing*:

The Elements of a Hypothesis Test

H_o: Null Hypothesis. A hypothesis about the distribution of the data (e.g. $U_1, U_2, \ldots, U_n \sim \ iid\ U(0,1)$)

T: Test Statistic

- T is a function of the data

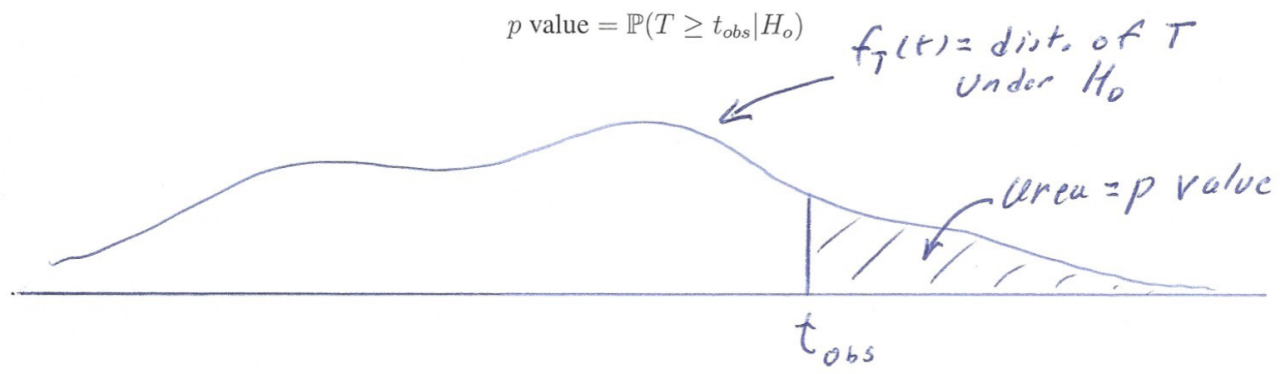
$$T = T(U_1, U_2, \ldots, U_n) = T(U_{1:n})$$

- We can evaluate $\mathbb{P}(T \in A | H_o)$, i.e. know distribution of T under H_o
- Expect T to be too large or too small under a natural alternative hypothesis ("H_a")
- T and H_a are supposed to be chosen before looking at the data! (Choosing T is as much art as science.)

t_{obs} : Observed value of T, $t_{obs} = T(U_{1:n})$

p value: If, for example, H_a suggests T will be large, as compared to what we would expect under H_o, then

$$p \text{ value} = \mathbb{P}(T \geq t_{obs} | H_o)$$



e.g. $n = 100$ and $T = \sum_{k=1}^{100} \mathbb{1}_{U_k > 0.9} = \#\{k : U_k > 0.9\}$. Then under H_o, T is Binomial$(100, 0.1)$. If, for example, we observe $t_{obs} = 18$ then

$$\begin{aligned} p \text{ value} &= \mathbb{P}(T \geq 18 | H_o) \\ &= 1 - \mathbb{P}(T \leq 17 | H_o) \\ &= \text{(Matlab) } 1 - \text{binocdf}(17, 100, 0.1) \\ &\cong 0.01 \end{aligned}$$

"We reject H_o in favor of H_a ($p = 0.01$)." (Would only expect this large, or larger, a deviation to happen at most 1% of the time if H_o were true.)

Remark: Actually, the alternative hypothesis could suggest other kinds of regions for T, not just too large or too small.

Obviously, there are many possiblbe tests...

Important Aside: On Testing Multiple Hypotheses

Suppose we perform many tests (say, M tests) with many different alternative hypotheses ($H_a^1, H_a^2, \ldots, H_a^M$), possibly with many different test statistics. And suppose that these M tests yield p values $\alpha_1, \alpha_2, \ldots, \alpha_M$, respectively. If we assume that H_o is true, then how surprised are we?

For every $\alpha \in (0, 1)$ and every test $k \in \{1, 2, \ldots, M\}$, let

$$E_k^\alpha = \{U_{1:n} : \text{reject } H_o \text{ in favor of } H_a^k, \text{ at } p \text{ value } \alpha\}$$

and let $\alpha_o = \min\{\alpha_1, \alpha_2, \ldots, \alpha_M\}$.

Then

$$\mathbb{P}\{\text{at least one test has } p \text{ value} \leq \alpha_o | H_o\} = \mathbb{P}\{\bigcup_{k=1}^{M} E_k^{\alpha_o} | H_o\}$$

$$\leq \sum_{k=1}^{M} \mathbb{P}(E_k^{\alpha_o} | H_o) = \sum_{k=1}^{M} \alpha_o = M\alpha_o = M \cdot (\text{minimum } p \text{ value})$$

e.g. $n = 100$, $T = \#\{k : U_k > 0.9\}$, and $t_{obs} = 4$. Suppose that under H_a^1, T is expected to be bigger than under H_o, and that under H_a^2, T is expected to be smaller. Then

$$\begin{aligned} \alpha_1 &= \mathbb{P}(T \geq 4 | H_o) \\ &= 1 - \text{binocdf(3,100,0.1)} \\ &= 0.992 \end{aligned}$$

and

$$\begin{aligned} \alpha_2 &= \mathbb{P}(T \leq 4 | H_o) \\ &= \text{binocdf(4,100,0.1)} \\ &= 0.0237 \end{aligned}$$

Putting these together: p value for the pair of tests $= 2 \times 0.0237 = 0.0474$.

2.2 From Uniform RV's to Arbitrary RV's Using the Inverse CDF

Suppose we have a really good PRNG, that generates (pseudo) rand variables that are nearly $iid\ U(0,1)$. How do we make a PRNG for other distributions? There are many methods for generating an iid sequence of an arbitrary random variable, X, given an iid sequence from $U(0,1)$.

Using the inverse cumulative distribution function

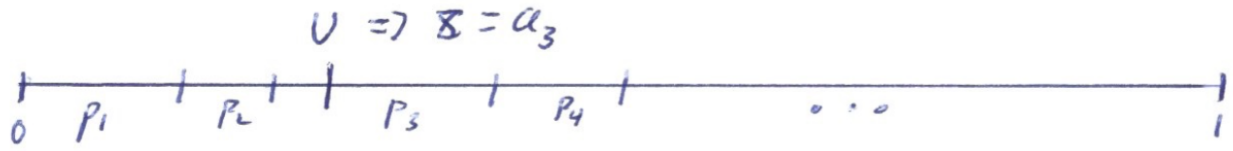
In general, suppose that X is a random variable with cdf $F_X(x)$ (recall that $F_X(x) \triangleq \mathbb{P}(X \leq x)$).

Discrete random variables

We are given

- $U \sim U(0,1)$
- $\{a_1, a_2, \ldots\}$ and $\{p_1, p_2, \ldots\}$ (possibly infinite sequences), where $a_1 < a_2 < \cdots$, $\sum_{k\geq 1} p_k = 1$, and $p_k > 0\ \forall k = 1, 2, \ldots$.

We want to generate X from U, with the result that $\mathbb{P}(X = a_k) = p_k\ \forall k = 1, 2, \ldots$.



In words, see where U falls and assign $X = a_k$ if U falls in the $k'th$ interval, in which case

$$\mathbb{P}(X = a_k) = \mathbb{P}(p_1 + \cdots p_{k-1} < U \leq p_1 + \cdots p_k) = p_k$$

(where, for convenience of notation, we define $p_0 = 0$). Formally,

$$X = \sum_{k\geq 1} a_k \mathbb{1}_{U \in (p_1 + \cdots p_{k-1}, p_1 + \cdots p_k]}$$

Remark: We can assign $X = k$ instead of a_k and then later change k to a_k.

Clumsy Matlab code for generating iid random variables from a discrete distribution

```
function X = drand(p,m,n)

% Use the Matlab editor to open a file.  Cut and paste this pdf into the
% editor and save it as 'drand.m' .  Then, from the command window, you can
% run drand.m using, for example,
%
% >> p=[.1 .2 .6 .05 .05];
% >> X=drand(p,20,2);
%
% In general, p is an S-length nonnegative vector that sums to 1.
%
% X is an m x n matrix of iid random observations
% taking values in 1,2,...,S with P(X(i,j)=k) = p(k).
%

% This function can take 1, 2, or 3 arguments ( drand(p), drand(p,m), or
% drand(p,m,n) ).
if nargin < 3, n = 1; end
if nargin < 2, m = 1; end

% get the size of the state space for X: 1,2,...,S
S = length(p);

% initialize X
X = zeros(m,n);

% loop over columns in X
for j = 1:n
    % loop over rows in X
    for i = 1:m
        % Determine X(i,j)
        % get a U(0,1) rv
        U = rand;
        %---- find out where it lands in the cdf of X ----%
        % initialize the cdf
        F = 0;
        % loop over possible values for X
        for k = 1:S
            % update the cdf
            F = F + p(k);
            % if U lands in this piece of the cdf, then set X(i,j)=k and
            % stop
            if U <= F
                X(i,j) = k;
                break
            end
        end
    end % loop over i
end % loop over j
```

Continuous random variables

Suppose now that we are given a strictly increasing and continuous function $F(x)$, that goes from zero to one on some specified interval. We want to produce a random variable X with cdf F, using a random variable $U \sim U(0,1)$.

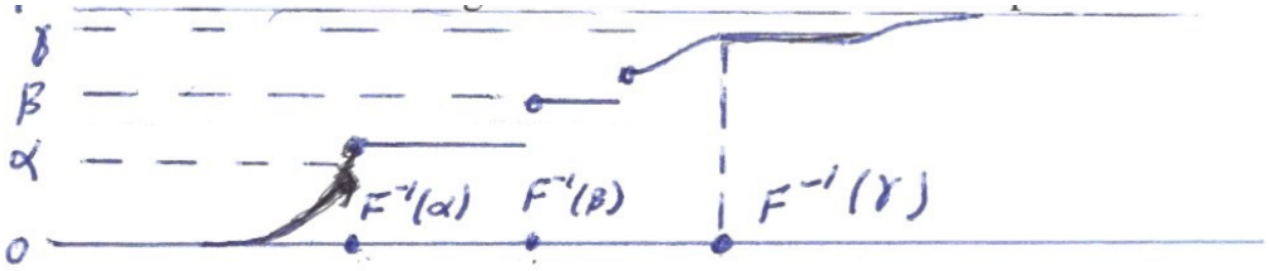
Consider $X = F^{-1}(U)$:

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$$

and, consequently, X has the desired cdf F.

Remarks

- More generally, F does not have to be continuous nor strictly increasing. We just need to be careful about defining F^{-1}: $F^{-1}(u) \triangleq \inf\{x : F(x) \geq u\}$, where the 'inf' of a set is the largest lower bound on all of the elements of the set. The figure below (which definitely needs improving!) includes some examples.



- With this definition of F^{-1}, there is no need to distinguish the discrete from the continuous case; in both cases, $X = F^{-1}(U)$.

e.g. Consider the continuous random variable with density function $f_X(x) = \alpha e^{-\alpha x}\mathbb{1}_{x \geq 0}$, for some $\alpha > 0$. This is the exponential distribution. To generate samples, first find $F_X(x)$:

$$F_X(x) = \int_{-\infty}^{x} \alpha e^{-\alpha t}\mathbb{1}_{t \geq 0} dt = \int_0^x \alpha e^{-\alpha t} dt = 1 - e^{-\alpha x}$$

Now solve for $F^{-1}(u)$, where $u \in (0,1)$:

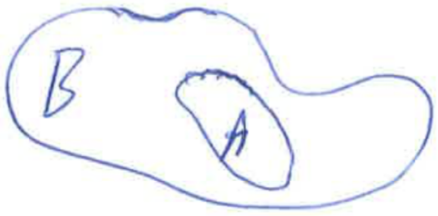
$$u = 1 - e^{-\alpha x} \Rightarrow e^{-\alpha x} = 1 - u \Rightarrow x = \frac{-1}{\alpha}\ln(1-u)$$

Finally, given $U \sim U(0,1)$, set $X = \frac{-1}{\alpha}\ln(1-U)$. (Notice that we could just as well have used $X = \frac{-1}{\alpha}\ln(U)$, since U and $1-U$ both have uniform distribution on $(0,1)$.)

But we need other methods if F^{-1} does not have a simple form...

Using rejection sampling (optional)

We start by generalizing the definition of a uniform random variable from the particular case, $U(01)$, to an aribtrary region in the plane:

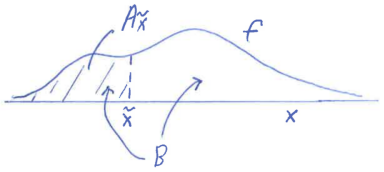


Defnition. Let B be a region in the plane ($B \subseteq \mathbb{R}^2$). We say that a pair of random variables, X, Y, have uniform distribution on B, $(X,Y) \sim U(B)$, if for every $A \subseteq B$

$$\mathbb{P}\big((X,Y) \in A\big) = \frac{\text{area}(A)}{\text{area}(B)}$$

Remark: So $(X,Y) \sim U(B) \Leftrightarrow f_{X,Y}(x,y) = \dfrac{\mathbb{1}_{(x,y)\in B}}{\text{area}(B)}$

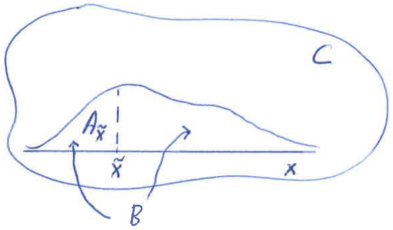
Now suppose that X is a continuous random variable with pdf f and cdf F. In many cases F is complicated and no explicit inverse is available. One way to sample X is to first sample from the uniform distribution on the area under f, and then keep the x-coordinate of the sample. Formally, let $B = \{(x,y) : 0 \leq y \leq f(x)\}$ and let (X,Y) be a sample from $U(B)$: $(X,Y) \sim U(B)$. To see that $X \sim F$, fix $\tilde{x} \in \mathbb{R}$ and define $A_{\tilde{x}} = \{(x,y) \in B : x < \tilde{x}\}$. Then



$$\mathbb{P}(X \leq \tilde{x}) = \mathbb{P}\big((X,Y) \in A_{\tilde{x}}\big) = \frac{\text{area}(A_{\tilde{x}})}{\text{area}(B)} = \text{area}(A_{\tilde{x}}) = \int_{-\infty}^{\tilde{x}} f(x)dx = F(\tilde{x})$$

But how are we to sample from $U(B)$? This is where *rejection sampling* comes in. Suppose that we can find a set $C \subseteq \mathbb{R}$ such that

- $B \subseteq C$, and
- It is easy to sample from $U(C)$



Then consider the following "rejection sampling" algorithm:

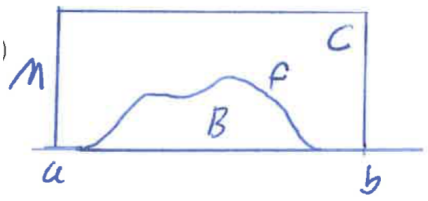
1. Choose $(X,Y) \sim C$ until $(X,Y) \in B$
2. Return X

Compute the distribution of the returned random variable, X:

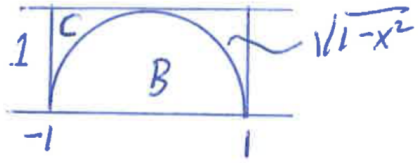
$$\mathbb{P}(X \leq \tilde{x}) = \mathbb{P}(X \leq \tilde{x} | (X,Y) \in B) = \frac{\mathbb{P}\big((X,Y) \in A_{\tilde{x}} \cap B\big)/\mathbb{P}\big((X,Y) \in C\big)}{\mathbb{P}\big((X,Y) \in B\big)/\mathbb{P}\big((X,Y) \in C\big)}$$

$$= \frac{\text{area}(A_{\tilde{x}} \cap B)}{\text{area}(B)} = \text{area}(A_{\tilde{x}}) = F(\tilde{x})$$

e.g. 1 Suppose $f(x) = 0$ unless $x \in [a, b]$, and $f(x) \leq M \ \forall x$. Take C to be the rectangle $[a, b] \times [0, M]$. To sample from $U(C)$, use $U_1, U_2 \sim \ iid \ U(0, 1)$, and then
$(X, Y) = (a + (b - a)U_1, MU_2) \sim U(C)$



e.g. 2 $f(x) = \frac{1}{Z}\sqrt{1 - x^2}\mathbb{1}_{-1 \leq x \leq 1}$. ($Z$ is the constant that "normalizes f," meaning we choose Z to make the area under f equal to one.) Now choose C to be a rectangle, as in the previous example, except that here $a = -1$, $b = 1$ and $M = 1$. Rejection sampling still works, even though we never calculated Z, or even took Z into account.



(To see that rejections sampling still works, just replace M by M/Z and $\sqrt{1 - x^2}$ by $\sqrt{1 - x^2}/Z$. The $Z's$ cancel.)

Remarks:

1. In this case, it is easy to compute Z: $Z = \int_{-1}^{1} \sqrt{1 - x^2} dx = \pi/2$. But it is not always so easy to compute the normalizing constant, as we shall see later in the semester.

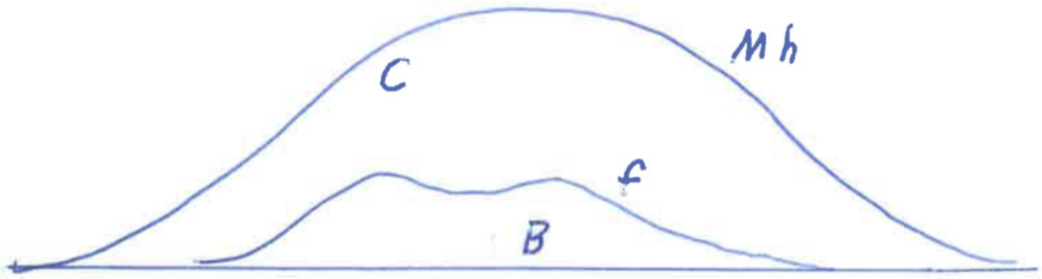
2. The expected fraction of iterations of the rejection-sampling algorithm that produce a random variable (i.e. the expected fraction that are "accepted") is just

$$\frac{\text{area}(B)}{\text{area}(C)} = \text{area}(B)/2$$

   which means that we could have estimated the area of B, and hence Z, from $2 \cdot$ (fraction accepted). This is an example of *stochastic approximation*, which we will study, more generally, in §II.

e.g. 3 Given a target density f, suppose that we can find a density h such that

1. we can draw samples from h (e.g. using the cdf method), and
2. for some M, $f(x) \leq M \cdot h(x) \ \forall x$



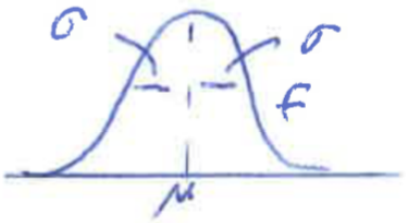
Now let $X \sim h$ and, given X, select $Y \sim U(0, Mh(X))$. Then $(X, Y) \sim U(C)$ and we can again do rejection sampling. To see that $(X, Y) \sim U(C)$:

$$f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y|x)$$
$$= h(x) \frac{1}{Mh(x)} \mathbb{1}_{0<y<Mh(x)} = \frac{1}{M} \mathbb{1}_{(x,y)\in C}$$
$$\Rightarrow \text{uniform on } C$$

Special case: Box-Muller method for generating Gaussian randomvariables (optional)

Finally, consider a Gaussian random variable with mean μ and standard deviation σ: $X \sim N(\mu, \sigma^2)$. The probability density function is

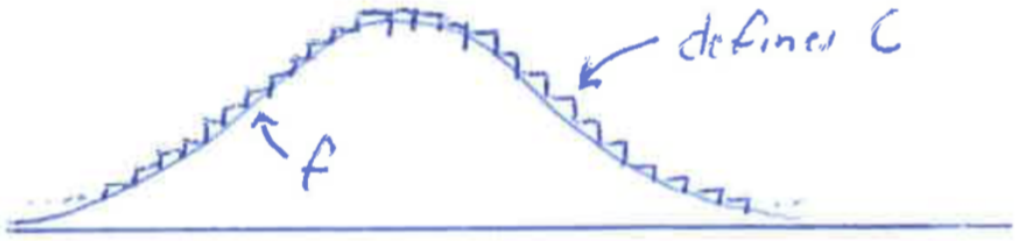
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



A useful fact is that if $Z \sim N(0, 1)$ ("standard normal"), then $\sigma Z + \mu \sim N(\mu, \sigma^2)$. Hence it is enough to sample from $N(0, 1)$, and then rescale by σ and relocate by μ. Because of this, we can assume, going forward, that $\mu = 0$ and $\sigma = 1$.

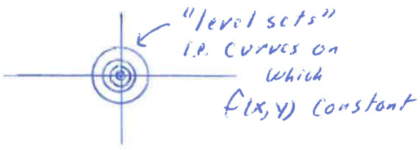
Unfortunately, whether or not $\mu = 0$ and $\sigma = 1$, both the cdf, $F(x)$, and its inverse are impossible to compute exactly, and approximations are computationally expensive.

One modern approach is based on rejection sampling, using a piecewise constant function that is everywhere bigger than or equal to f, but also tightly approximates f. This defines a region C for rejection sampling, and look-up tables make for fast code for sampling from $U(C)$.



An older method, but one that is fast and easy to program, and has the virtue of affording some additional insight into Gaussian random variables, is to use the Box-Muller transform. Recall that any point $(x, y) \in \mathbb{R}^2$ can be represented in its "polar coordinates" (r, θ), where $r = \sqrt{x^2 + y^2}$ and $\theta = \arctan(y/x)$. Since $x = r\cos(\theta)$ and $y = r\sin(\theta)$, we have the change-of-variables formula $dxdy = rdrd\theta$. In other words, if $g(x, y)$ is any function of (x, y) then integration of $g(x, y)dxdy$ can be performed in polar coordinates by integrating $g(r\cos(\theta), r\sin(\theta))rdrd\theta$.

The idea of the Box-Muller method is to use two independent uniform random variables, $U_1, U_2 \sim \ iid\ U(0,1)$, to get two independent standard normal random variables, X, Y, as follows. Let $X, Y \sim \ iid\ N(0,1)$, in which case



$$\begin{aligned} f_{X,Y}(x,y)dxdy &= f_X(x)f_Y(y)dxdy \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dxdy \\ &= \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}dxdy \quad \text{Remark: the distribution is } \textit{circularly symmetric}! \\ &= \frac{1}{2\pi}e^{-\frac{r^2}{2}}rdrd\theta \\ &= \left(\frac{d\theta}{2\pi}\right)\left(re^{-\frac{r^2}{2}}dr\right) \\ &= h(\theta)k(r)d\theta dr \end{aligned}$$

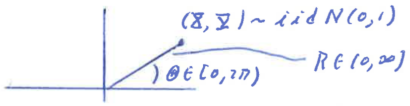
Where $h(\theta) \triangleq 1/(2\pi)$ and $k(r) \triangleq re^{-r^2/2}$, for all $\theta \in [0, 2\pi]$ and $r \geq 0$. This means that if we can generate $\Theta \sim h$ and $R \sim k$ then we can generate $X, Y \sim \ iid\ N(0,1)$ by setting $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$.

Now let U_1 and U_2 be two independent $U(0,1)$ random variables. It's easy to sample from h: just set $\Theta = 2\pi U_1$. As for k, use the inverse cumulative distribution:

$$k(r) = re^{-\frac{r^2}{2}}\mathbb{1}_{r\geq 0} \Rightarrow K(r) \triangleq \int_0^r se^{-\frac{s^2}{2}}ds = \cdots = 1 - e^{-\frac{r^2}{2}}$$

Solving $u = 1 - e^{-\frac{r^2}{2}}$ gives $K^{-1}(u) = \sqrt{-2\log(1-u)}$ which means that $R = \sqrt{-2\log(1-U_2)}$ has the required distribution. Or we can simply use $R = \sqrt{-2\log U_2}$, since U_2 has the same distribution as $1 - U_2$.

In summary, we generate a pair (X, Y) of independent $N(0,1)$ random variables from a pair (U_1, U_2) of independent $U(0,1)$ random variables using



$$(X, Y) = \left(\sqrt{-2\log U_2}\cos(2\pi U_1), \sqrt{-2\log U_2}\sin(2\pi U_1)\right)$$

As a final remark, and another application of polar coordinates, here's a cool way to compute the normalization constant (i.e. $\sqrt{2\pi}$) for the standard normal.

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx &= \Big(\int_{x=-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \int_{y=-\infty}^{\infty} e^{-\frac{y^2}{2}} dy\Big)^{1/2} \\
&= \Big(\int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}} dydx\Big)^{1/2} \\
&= \Big(\int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dydx\Big)^{1/2} \\
&= \Big(\int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} e^{-\frac{r^2}{2}} r d\theta dr\Big)^{1/2} \\
&= \Big(2\pi \int_{r=0}^{\infty} r e^{-\frac{r^2}{2}} d\theta dr\Big)^{1/2} \\
&= \Big(2\pi(-e^{-\frac{r^2}{2}})|_0^{\infty}\Big)^{1/2} \\
&= \sqrt{2\pi}
\end{aligned}$$

2.3 Monte Carlo Integration

2.3.1 A Nifty Application of the LLN

e.g. Consider

$$I = \int_0^1 u^5 du = \frac{1}{6} u^6|_0^1 = \frac{1}{6}$$

Note that if $U_{1:n} \sim \ iid\, U(0,1)$ then

$$\frac{1}{n}\sum_{k=1}^{n} U_k^5 \xrightarrow[n\to\infty]{LLN} E[U^5] = \int_{-\infty}^{\infty} u^5 \mathbb{1}_{u\in(0,1)} du = \int_0^1 u^5 du = I$$

Or, using Matlab, repeat the following code many times

```
>> n=10^6;
>> u=rand(n,1).^5;
>> mean(u)
```

to get .1667, .1667, .1666, ... This is an example of *Monte Carlo Integration*.

e.g. Now consider a more difficult integral: $I = \int_0^1 h(u)du$, where

$$h(u) = \cos(5\log u)\log(1+u)du$$

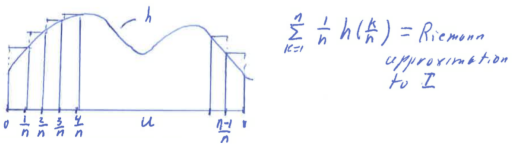
This is hard to evaluate exactly, but just as easy as the previous example using the LLN: repeat the following code

```
>> n=10^6;
>> u=rand(n,1);
>> h=cos(5*log(u)).*log(1+u);
>> mean(h)
```

this time, to get .0437, .0433, .0438, ... as approximations for I.

What sorts of precise statements can we make about accuracy? As we shall see shortly, these are addressed with the usual tools of statistical estimation, including confidence intervals and central limit theorem approximations.

The results so far are not impressive, since we could do just as well, if not better and more simply, by using the Riemann approximation.



Almost the identical code will give us the Riemann approximation:

```
>> n=10^6;
>> u=(1:n)/n;
>> h=cos(5*log(u)).*log(1+u);
>> mean(h)
```

which produces the estimate .0433. Furthermore, this approximation comes with guarantees in terms of rate of convergence, bounds on the error, and a host of other well-studied "numerical methods" for improving accuracy.

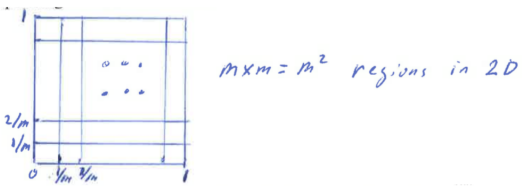
On the other hand, these deterministic numerical methods do not easily extend to higher dimensions. Consider this example:

e.g.

$$I = \int_{u_1=0}^{1} \cdots \int_{u_{100}=0}^{1} (\sum_{k=1}^{100} k u_k)^2 du_1 \cdots du_{100}$$

which, with a little work, can be done exactly to get $I = 6.4038208\bar{3} \times 10^6$.

The straightforward generalization of the Riemann approach requires, first, a discretization of the coordinates. Let's say that we use m points for each coordinate. If the integral were two dimensional (instead of 100 dimensional) then we would end up with m^2 squares (instead of m intervals, as in one dimension) and m^2 corresponding function evaluations.



Since this integral is one-hundred dimensional, we would require m^{100} function evaluations. Even if we were to use only two intervals per coordinate ($m = 2$) we would require an unworkable $2^{100} > 10^{30}$ function evaluations.

As for the Monte Carlo approach, let $U(1), \ldots, U(n) \sim \ iid\ U(0,1)^{100}$ (so $U(k) = (U_1(k), \ldots, U_{100}(k))$, where $U_1(k), \ldots, U_{100}(k) \sim \ iid\ U(0,1)$ for every $k = 1, 2, \ldots, 100$). Now let $h(u_{1:100}) = (\sum_{k=1}^{100} k u_k)^2$

and observe that

$$\frac{1}{n}\sum_{k=1}^{n} h(U(k)) \xrightarrow[n\to\infty]{LLN} E[h(U_{1:100})] = \int_{u_1=0}^{1} \cdots \int_{u_{100}=0}^{1} (\sum_{k=1}^{100} k u_k)^2 du_1 \cdots du_{100} = I$$

In Matlab, repeating

```
>> n=10^5;
>> u=rand(n,100);
>> h=(u*(1:100)').^2;
>> mean(h)
```

yields $6.405 \times 10^6,\ 6.403 \times 10^6,\ 6.403 \times 10^6$, etc.

But Monte Carlo Integration is not without its difficulties (especially in high dimensions, which we will come back to a little later).

e.g. Estimate $I = \int_0^1 h(x)dx$, this time with

$$h(x) = |cos(1000\pi x)|e^{-100x}$$

(For no particular reason, I've switched the 'dummy variable' from u to x.) I ran the code:

```
n=500;
x=rand(n,1);
h=abs(cos(1000*pi*x)).*exp(-100*x);
I=mean(h);
```

four times, and got: 0.0063, 0.0031, 0.0080, and 0.0067, which are kind of all over the place.

Can we trust these numbers? They seem to be hovering around 0.0055, maybe. But take a close look at the integral: e^{-100x} will typically be quite small. (In fact, you can verify that $E[e^{-100X}] \approx 0.01$ for $X \sim U(0,1)$.) Since $|cos(1000\pi x)| \leq 1$ for all $x \in (0,1)$, we are mostly sampling values that are quite small (due to the exponential term), and largely missing all the "action," which is happening near zero where the exponential term is not so small and the rapid fluctuations of $|cos(1000\pi x)| \leq 1$ are largely determining the value of I.

It's tempting to try to somehow bias the samples so that they are more typically near zero, rather than being uniform on the interval. But then our simple justification, via the LLN, disappears. This is where "importance sampling" comes in: by a surprisingly simple trick we can indeed bias the samples so that they are "near the action," and, if we do it carefully, then we still have the guarantee of converging (as $n \to \infty$) to the true value of the integral.

Here then is the simple (but indeed very clever) approach known as *importance sampling*:

2.3.2 Importance Sampling

The idea here is to explore different sampling distributions, beyond just the uniform distribution. The goal is the same: estimate the value of a given interval.

In general, let $h : \mathbb{R}^d \to \mathbb{R}$, and write $h(x)$ instead of $h(\vec{x})$ or $h(x_{1:d})$ so as to keep the notation simple. We are interested in

$$I = \int_{\mathbb{R}^d} h(x)dx$$

Instead of sampling from $U(0,1)^d$, we will sample, more generally, from some pdf f on $\mathbb{R}^d$. Given $X(1), \ldots, X(n) \sim \ iid\ f$, consider the approximation

$$\hat{I}_n \triangleq \frac{1}{n}\sum_{k=1}^{n} \frac{h(X(k))}{f(X(k))} \xrightarrow[n\to\infty]{LLN} E[\frac{h(X)}{f(X)}] = \int_{\mathbb{R}^d} \frac{h(x)}{f(x)} f(x)dx = I$$

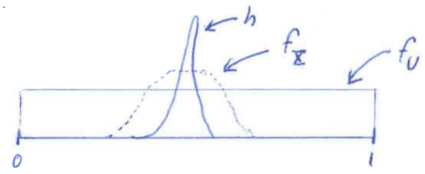
That's amazing. We can sample from just about *any* distribution and still recover the same integral! Now, from our new and more general perspective, we can look back on the earlier approach as a special case:

$$I = \int_{R^1} g(x)\mathbb{1}_{x\in(0,1)}dx$$

for some function g, $h(x) = g(x)\mathbb{1}_{x\in(0,1)}$, and $f(x) = 1$ for all $x \in (0,1)$.

We can't actually use *any* sampling distribution since, at the least, it had better have some "mass" wherever h happens to have mass. Otherwise, we would miss entirely some of the values of h—namely the ones where $h(x) \neq 0$ but $f(x) = 0$. In other words, we require that $\{x : h(x) \neq 0\} \subseteq \{x : f(x) > 0\}$ or, what is the same thing, $f(x) = 0 \Rightarrow h(x) = 0$.

The key to making this work is to choose f to be big where there is action in h (hence the name importance sampling). In the attached image, for example, f_X would be a much better choice than f_U.



Let's return to our previous example, with a slight change in notation: $I = \int_{-\infty}^{\infty} h(x)dx$, and

$$h(x) = |cos(1000\pi x)|e^{-100x}\mathbb{1}_{x\in(0,1)}$$

This is just begging for an exponential sampling distribution, namely $X \sim \exp(.01)$, which means the exponential density with mean 0.01:

$$f(x) = 100e^{-100x}$$

which certainly fits the bill—big when h is big and small when h is small.

Given $X(1), \ldots, X(n) \sim \text{exp}(.01)$,

$$\begin{aligned}
\hat{I}_n &= \frac{1}{n}\sum_{k=1}^{n} \frac{|cos(1000\pi X(k))|e^{-100X(k)}\mathbb{1}_{X(k)\in(0,1)}}{100e^{-100X(k)}} \\
&= \frac{1}{n}\sum_{k=1}^{n} \frac{1}{100}|cos(1000\pi X(k))|\mathbb{1}_{X(k)\in(0,1)} \\
&\xrightarrow[n\to\infty]{LLN} E_f[\frac{1}{100}|cos(1000\pi X)|\mathbb{1}_{X\in(0,1)}] \quad \text{(I used } E_f \text{ to indicate that } X \text{ has } f \text{ as its pdf)} \\
&= \int_0^1 \frac{1}{100}|cos(1000\pi X)|100e^{-100x}dx \\
&= \int_0^1 |cos(1000\pi X)|e^{-100x}dx = I
\end{aligned}$$

Let's try it:

```
n=500;
x=exprnd(0.01,n,1);
h=(x>0).*(x<1).*abs(cos(1000*pi*x)).*exp(-100*x); % "(x>0).*(x<1).*"
                    % because the integral is from x=0 to x=1.
hOVERf=(x<1).*abs(cos(1000*pi*x))/100; % since f=100*exp(-100*x).*(x>0)
I=mean(hOVERf);
```

Four runs came out with 0.0062, 0.0063, 0.0064, and 0.0066. Looks better than 0.0063, 0.0031, 0.0080, and 0.0067, which is what we got using the uniform sampling distribution. And "looks better" is good, but we will definitely want to make more precise statements, e.g. using *confidence intervals*. Remember those?

2.3.3 Estimating an Integral—with confidence

As far as the statistics world goes, we are in familiar territory. We have a target parameter, I, and an estimator, $\hat{I}_n$. Suppose that the estimator is *unbiased* (i.e. $E[\hat{I}_n] = I$), and we happen to know $\text{Var}(\hat{I}_n)$ (the variance of the estimator). Then, if $\hat{I}_n$ is normally distributed ($\hat{I}_n \sim N(I, \text{Var}(\hat{I}_n))$, we can reason as follows to get an *interval estimate* (aka *confidence interval*) for I: Let $C \in (0, 1)$ ("confidence level") and

let Z be a standard normal random variable, $Z \sim N(0,1)$. Choose z_o such that $F_Z(z_o) = \frac{C+1}{2}$. Then[4]

$$\begin{aligned}\mathbb{P}\big(-z_o < \frac{\hat{I}_n - I}{\sqrt{\text{Var}(\hat{I}_n)}} < z_o\big) &= \mathbb{P}(-z_o < Z < z_o) \\ &= F_Z(z_o) - F_Z(-z_o) \\ &= F_Z(z_o) - (1 - F_Z(z_o)) \\ &= 2F_Z(z_o) - 1 = C\end{aligned}$$

So

$$C = \mathbb{P}\big(-z_o < \frac{\hat{I}_n - I}{\sqrt{\text{Var}(\hat{I}_n)}} < z_o\big) = \mathbb{P}\left(\hat{I}_n - z_o\sqrt{\text{Var}(\hat{I}_n)} < I < \hat{I}_n + z_o\sqrt{\text{Var}(\hat{I}_n)}\right)$$

In other words, the *random interval* $\left(\hat{I}_n - z_o\sqrt{\text{Var}(\hat{I}_n)}, \hat{I}_n + z_o\sqrt{\text{Var}(\hat{I}_n)}\right)$ contains I with probability C. Probability, here, is with respect to the random sequence $X(1), X(2), \ldots, X(n)$ from which we constructed $\hat{I}_n$ in the first place.

We can make this a bit more explicit by remembering where $\hat{I}_n$ came from, in the first place:

$$\hat{I}_n = \frac{1}{n}\sum_{k=1}^{n}\frac{h(X(k))}{f(X(k))}$$

where $X(1), \ldots, X(n) \sim iid\ f(x)$. In this case

$$E[\hat{I}_n] = \frac{1}{n}\sum_{k=1}^{n} E\big[\frac{h(X(k))}{f(X(k))}\big] = \frac{1}{n}\sum_{k=1}^{n}\int_{\mathbb{R}^d}\frac{h(x)}{f(x)}f(x)dx = \frac{1}{n}\sum_{k=1}^{n} I = I$$

(So $\hat{I}_n$ is indeed unbiased.) As for the $\text{Var}(\hat{I}_n)$,

$$\begin{aligned}\text{Var}(\hat{I}_n) &= \text{Var}\big(\frac{1}{n}\sum_{k=1}^{n}\frac{h(X(k))}{f(X(k))}\big) \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\text{Var}\big(\frac{h(X(k))}{f(X(k))}\big) \quad \text{(Rules of the Road!)} \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\text{Var}\big(\frac{h(X)}{f(X)}\big) \quad (X(1), \ldots, X(n) \text{ are iid, with a common distribution } X \sim f(x)) \\ &= \frac{1}{n}\text{Var}\big(\frac{h(X)}{f(X)}\big) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

[4] NB: If W is $N(\mu, \sigma^2)$ then $\frac{W-\mu}{\sigma} \sim N(0,1)$. In fact, if $W_1, \ldots, W_n$ are *independent*, with $W_k \sim N(\mu_k, \sigma_k^2),\ k = 1, \ldots, n$, then for any constants $a_1, \ldots, a_n$, $\sum_k a_k W_k \sim N(\sum_k a_k\mu_k, \sum_k a_k^2\sigma_k^2)$.

where $\sigma^2 = \text{Var}\big(\frac{h(X)}{f(X)}\big)$.

In summary, $\big(\hat{I}_n - z_o\frac{\sigma}{\sqrt{n}}, \hat{I}_n + z_o\frac{\sigma}{\sqrt{n}}\big)$ contains I with probability C.

Remarks

1. As we have already observed, what we started with (in §2.3.1) was a special case, with $I = \int_0^1 h(x)dx = \int_{-\infty}^{\infty} h(x)\mathbb{1}_{x\in(0,1)}dx$ and $X(1), \ldots, X(n) \sim iid\ U(0,1)$.

2. How good an approximation is $\hat{I}_n$ of I? As already noted, $E[\hat{I}_n] = I$. Hence, the *mean squared error*, $E[(\hat{I}_n - I)^2]$, is the same as the variance, which we calculated earlier: $E[(\hat{I}_n - I)^2] = \frac{\sigma^2}{n}$, where σ^2 is the variance of the ratio
$$\sigma^2 = \text{Var}\big(\frac{h(X)}{f(X)}\big)$$
when $X \sim\ f$. The key then is to make σ^2 small—i.e. choose f to be of roughly the same shape as h, so that the ratio, $h(x)/f(x)$, is roughly a constant. (But of course f first needs to be *useable*—we have to be able to sample iid sequences from f.)

3. But how big is σ? If it can not be computed directly then we can still estimate it using the same samples $X(1), \ldots, X(n)$:
$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}\big(\frac{h(X(k))}{f(X(k))} - \hat{I}_n\big)^2 \xrightarrow[n\to\infty]{LLN} \sigma^2 \tag{3}$$
In this case, the confidence interval becomes $\big(\hat{I}_n - z_o\frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{I}_n + z_o\frac{\hat{\sigma}_n}{\sqrt{n}}\big)$.

4. Recall that z_o is chosen to make $F_Z(z_o) = \frac{C+1}{2}$. By far the most typical value of C is C =0.95, in which case z_o is very nearly two. If, for example , $\hat{I}_n$ turns out to be 83.4 (say) and $\hat{\sigma}_n \approx 2.8$ then we might report that "83.4 $\pm$ 5.6 is an approximate 95% confidence interval for the value I of the integral."

5. Observe that the confidence interval shrinks rather slowly: we need four times as many samples to get half as wide an interval. Also, when using an estimated σ, the interval is only as reliable as the estimate of σ. This situation, in general, is part of the motivation leading to the Bootstrap, which we will be discussing a little later in this chapter.

Let's look at two examples, with confidence intervals:

e.g. 1 First, we will revisit the example $I = \int_{-\infty}^{\infty} h(x)dx$, with
$$h(x) = |cos(1000\pi x)|e^{-100x}\mathbb{1}_{x\in(0,1)}$$
Recall that sampling from $f_X(x) = \mathbb{1}_{x\in(0,1)}$, using 500 samples, gave us 0.0063, 0.0031, 0.0080, and 0.0067 over four independent runs. On the other hand, sampling from $f_X(x) = 100e^{-100}\mathbb{1}_{x>0}$, and adjusting $\hat{I}_n$ accordingly,
$$\hat{I}_n = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{100}|cos(1000\pi X(k))|\mathbb{1}_{X(k)\in(0,1)}$$
yielded 0.0062, 0.0063, 0.0064, and 0.0066, an apparent improvement.

It is easy to modify the code to include estimated confidence intervals:

uniform sampling

```
n=500;
x=rand(n,1);
h=abs(cos(1000*pi*x)).*exp(-100*x);
I=mean(h);
SigmaHat=sqrt(mean((h-I).^2));
conf=I+(SigmaHat/sqrt(n))*[-2,2];
size=conf(2)-conf(1);
disp(['Estimate of I: ',num2str(I)])
S=['Approximate 95 prcent CI:','(',num2str(conf(1)),',',num2str(conf(2)),')'];
disp(S)
disp(['Size of CI: ',num2str(size)])
```

A single run yielded

```
Estimate of I: 0.0040611
Approximate 95 percent CI: (0.0010109,0.0071113)
Size of CI: 0.0061004
```

The confidence interval is wider than the estimate!

importance sampling

```
n=500;
x=exprnd(0.01,n,1);
h=(x>0).*(x<1).*abs(cos(1000*pi*x)).*exp(-100*x); % "(x>0).*(x<1).*"
                    % because the integral is from x=0 to x=1.
hOVERf=(x<1).*abs(cos(1000*pi*x))/100; % since f=100*exp(-100*x).*(x>0)
I=mean(hOVERf);
Sigma=sqrt(mean((hOVERf-I).^2));
conf=I+(Sigma/sqrt(n))*[-2,2];
size=conf(2)-conf(1);
disp(['Estimate of I: ',num2str(I)])
S=['Approximate 95 percent CI:','(',num2str(conf(1)),',',num2str(conf(2)),')'];
disp(S)
disp(['Size of CI: ',num2str(size)])
```

```
Estimate of I: 0.0064522
Approximate 95 percent CI:(0.0061878,0.0067167)
Size of CI: 0.0005289
```

e.g. 2 (cold and lonely in high dimensions) The volume (aka length) of the "unit ball" in $\mathbb{R}^1$ is 2:

$$\int_{x\in\mathbb{R}^1} \mathbb{1}_{x^2<1} dx = 2$$

In two dimensions the volume (aka area) is π:

$$\int_{x_1\in\mathbb{R}^1}\int_{x_2\in\mathbb{R}^1} \mathbb{1}_{x_1^2+x_2^2<1} dx = \pi$$

More generally, for any integer $d > 0$, let $B_d = \{(x_1,\ldots,x_d) : \sum_k x_k^2 < 1\}$ be the d-dimensional unit ball in $\mathbb{R}^d$. If d is even, then

$$\text{Vol}(B_d) = \int_{x\in\mathbb{R}^d} \mathbb{1}_{x\in B_d} dx = \int_{x_1\in\mathbb{R}^1}\cdots\int_{x_d\in\mathbb{R}^1} \mathbb{1}_{\sum_{k=1}^d x_k^2<1} dx_d\cdots dx_1 = \frac{\pi^{d/2}}{\frac{d}{2}!}$$

(There is a similar, but messier, formula for those cases when d is odd.) Let's pretend like we don't know these formulas, haven't a clue about how to do the integrals exactly, and have no recourse but to try to estimate the integrals. This may not have happened to you lately, but this kind of thing happens all the time.

What should we use for the sampling distribution? Here's an easy choice: $X \in U(-1,1)^d$, i.e. sample from the uniform distribution on the cube with sides $x_k \in (-1,1)$ for every $k = 1,\ldots,d$. The cube is easy (the coordinates are just iid uniform, each on $(-1,1)$), fast, and "tight"—meaning that it touches the ball at each of the 2^d points $\{(x_1,\ldots,x_d) : x_k = 0$ for all but one value of k, at which $x_k \in \{-1,1\}\}$. The prescription, then, is to let $h(x) = \mathbb{1}_{x\in B_d}$ and $f = 2^{-d}\mathbb{1}_{x\in(-1,1)^d}$. In which case, if $X(1),\ldots,X(n) \sim iid\ f$ then

$$\begin{aligned}\hat{I}_n &= \frac{1}{n}\sum_{i=1}^n \frac{h(X(i))}{f(X(i))} \\ &= \sum_{i=1}^n \frac{\mathbb{1}_{X(i)\in B_d}}{2^{-d}\mathbb{1}_{X(i)\in(-1,1)^d}} \\ &= \sum_{i=1}^n 2^d \mathbb{1}_{X(i)\in B_d}\end{aligned}$$

So in this case, importance sampling reduces to rejection sampling—just count the fraction of samples from f that land in B_d, but don't forget to multiply by the normalizing constant 2^d. As for confidence intervals, we can use equation (3),

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \Big(\frac{h(X(i))}{f(X(i))} - \hat{I}_n\Big)^2 = \frac{1}{n}\sum_{i=1}^n \big(2^d\mathbb{1}_{X(i)\in B_d} - \hat{I}_n\big)^2$$

to get $\hat{I}_n \pm 2\frac{\hat{\sigma}_n}{\sqrt{n}}$ as an approximate 95% CI for I.

Monte Carlo estimation of the volume of B_d
uniform sampling distribution

```
% The problem is to estimate the volume of the d-dimensional ball, with
% radius 1 (aka the "unit ball in d dimensions"). The approach is
% importance sampling, which can be viewed as a generalization of rejection
% sampling.

% This example uses the uniform sampling distribution, on (-1,1)^d

% dimensions    A vector--each component of which is a number, d, of
%               dimensions in which we will estimate the volume of the unit
%               ball. Each d assumed to be even, for convenience.

% nsams         The number of samples drawn from the sampling distribution,
%               which is the unifrom distribution on the d-dimensional
%               cube, with sides x_k in (-1,1), for k=1:d.

% nruns         Number of Monte Carlo runs (each using n samples)

dimensions=[2 4 8 16 32];
nsamps=1000000;             % 1,000,000
nruns=3;

ndims=numel(dimensions);
IHat=zeros(nruns,ndims); % Each row is a run and each column is
                                % a number, d,  of dimensions. The entries
                                % are estimated volumes.
CIR=zeros(nruns,ndims);  % Will contain the radiuses of the approximate
                         % 95% confidence intervals, using normal approx.
npos=zeros(nruns,ndims); % How many of the nsampls samples landed in the
                         % unit ball

% Compute the true volumes, using the exact formula for even number of
% dimensions.

HalfD=dimensions/2;
TrueVols=(pi.^HalfD)./factorial(HalfD);

for dim=1:ndims
    d=dimensions(dim);
    for run=1:nruns
        % Sample from the sampling distribution
        u=2*rand(nsamps,d)-1;   % an nxd array of independent numbers, each
                                % unifrom on (-1,1)
        % Compute h, which is just the indicator function on the unit ball
        h=sum(u.^2,2)<1;   % h is an nx1 (row) vector, with ones
                           % indicating samples that fall in the unit ball,
                           % and zero indicating those outside ball
        % Compute f
        f=2^(-d)*ones(nsamps,1);
        % Estimator is mean of h/f
        IHat(run,dim)=mean(h./f);
        % Compute the radiuses of the (approx.) 95% CI's
        SHat=sqrt(mean( (h./f - IHat(run,dim)).^2 )); % Estimated st. dev.
        CIR(run,dim)=2*SHat/sqrt(nsamps); % 2*sigma/sqrt(n)
        % A good measure of efficiency is number of the nsamps that landed
        % in the unit ball
        npos(run,dim)=sum(h);
    end
end
```

The code labeled "Monte Carlo estimation of the volume of B_d, uniform sampling distribution" produces the necessary quantities, as well as a few other quantities that will be of interest.

Here are the results of 3 Monte Carlo runs ($n = 10^6$) for each of dimensions d=2,4,8,16, and 32 (correct volumes are in the first row, in bold):

d	2	4	8	16	32
$\boldsymbol{Vol(B_d)}$	**3.1416**	**4.9348**	**4.0587**	**0.2353**	**4.3031**$\times 10^{-6}$
run 1	3.1381$\pm$ 0.0033	4.9352 $\pm$ 0.0148	4.0786 $\pm$ 0.0641	0.0655 $\pm$ 0.1311	0 $\pm$ 0
run 2	3.1399 $\pm$ 0.0033	4.9369 $\pm$ 0.0148	3.9900 $\pm$ 0.0634	0.1966 $\pm$ 0.2270	0 $\pm$ 0
run 3	3.1421 $\pm$ 0.0033	4.9339 $\pm$ 0.0148	4.0474 $\pm$ 0.0639	0.3277 $\pm$ 0.2931	0 $\pm$ 0

The results for dimensions 2, 4, and 8 are great! The confidence intervals are narrow, and each of them either includes or very nearly includes the correct value. But what is going wrong in dimensions 16 and 32? In particular, in dimension $d = 32$, how is it possible that every run keeps giving zero as the estimate of I? All we are doing is recording the fraction of the samples from $(-1, 1)^d$ that fall in B_d. Is the fraction really that small? Let's calculate: according to the table (first row) the volume of B_{32} is about $.2353\times 10^{-6}$, and since we are uniformly sampling from $(-1, 1)^{32}$, which has volume 2^{32}, the chances of a sample landing in B_{32} are about $(4.3031 \times 10^{-6})/2^{32} \approx 10^{-15}$! We will not be seeing anything but zeros anytime soon. As for $d = 16$, the corresponding ratio is $0.2353/2^{16} \approx 3.6 \cdot 10^{-6}$.

These calculations are consistent with the following table, which shows the number of the 1,000,000 samples that landed inside the ball B_d, for each dimension and each trial:

d	2	4	8	16	32
run 1	784516	308452	15932	1	0
run 2	784968	308557	15586	3	0
run 3	785534	308370	15810	5	0

In dimensions 16 and 32, we are getting very little out of a lot of computing. The sampling is not only inefficient, it is leading to wildly unreliable estimates.

It might seem odd that B_d, which touches the cube $(-1, 1)^d$ at $2d$ locations, would take up such a small fraction of the cube's volume. But consider that the distance from the origin to any corner of the cube (e.g. the point with all coordinates equal to 1) is $\sqrt{d}$, whereas B_d has radius equal to 1. There's a lot of room in those corners!

Evidently, in higher dimensions we will need a better sampling distribution. To get some intuition, observe that most of the volume of B_{32} is near the surface. To make this precise, let $B_{32}(r)$ represent the 32-dimensional ball with radius r and centered at the origin (so $B_{32} = B_{32}(1)$). A universal scaling law is the power law for volumes in dimension d: For any $A \in \mathbb{R}^d$ and any $a \in \mathbb{R}^1,\ a > 0$, let $rA = \{x : \frac{x}{a} \in A\}$. Then $\text{Vol}(rA) = r^d\text{Vol}(A)$, where $\text{Vol}(A)$ refers to the $d-$dimensional volume of A. (This is the basic reason that babies have trouble regulating their temperatures. The surface of a body scales like r^2, but the volume scales like r^3. Everything else being equal—including the shape!—the surface area to volume ratio gets larger as the body gets smaller, meaning that the relative rate of gain or loss of heat is higher in smaller bodies. Put simply, there is relatively more area through which to move heat.) Apply these scalings to balls centered at the origin:

$$\frac{\text{Vol}(B_d(0.9)}{\text{Vol}(B_c(1.0)} = (0.9)^d$$

When $d = 16,\ (0.9)^d$ is about 0.1853, which means that over 80% of the volume of B_{16} is within distance 0.1 of the surface. Using the same reasoning, you can check for yourself that when $d = 32$, over 96% of the volume is within 0.1 of the surface. To use some of the same language we used earlier, most of the action in I (when d is large) is near the surface of the ball. We will need to pick a sampling distribution

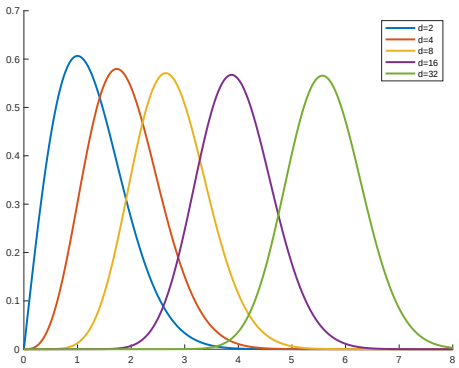
that emphasizes the region near to the surface of B_d.

For now, fix $d = 32$ and consider $Z_{1:32} \sim \ iid\ N(0,1)$ as a possible sampling distribution for the problem of computing the volume of B_{32}. This is called the 32-dimensional standard normal distribution. The joint pdf is just the product of 32 standard normal pdf's

$$f_{Z_{1:32}}(z_{1:32}) = \left(\frac{1}{\sqrt{2\pi}}\right)^{32} e^{-\frac{1}{2}\sum_1^{32} z_k^2} \tag{4}$$

Notice that f depends only on the distance to the origin, and hence it is circularly symmetric. This is a good place to start since B_{32} is also circularly symmetric. But in addition, we want a substantial amount of the mass of Z to lie near the surface of B_{32}, i.e. at distance 1 from the origin.

The joint distribution on $Z = Z_{1:32}$ (joint pdf—4) determines the distribution on the distance of a random sample of Z from the origin, i.e. $f_{Z_{1:32}}$ determines f_R, where $R \triangleq \sqrt{\sum_1^{32} Z_k^2}$. Indeed, it turns out that R has been well studied; it is known as the "Chi distribution with 32 degrees of freedom." It's shape, more generally as a function of d, can be seen in the accompanying figure. The mode (highest point) occurs at $R = \sqrt{31}$, or, more generally, at $\sqrt{d-1}$. So using iid standard normal random variables will generally put us much too far from the surface of B_{32}, which, after all, is only one unit from the origin.



Let's try, instead, $X_{1:32} \sim \ iid\ N(0, \sigma^2)$, and then adjust σ so that its mode is at or near 1:

$$R = \sqrt{\sum_1^{32} X_k^2} \sim \sqrt{\sum_1^{32} (\sigma Z_k)^2} = \sigma \sqrt{\sum_1^{32} Z_k^2}$$

which therefore has mode at $\sigma\sqrt{31}$. Setting this to 1, we get $\sigma = 1/\sqrt{31}$. More generally, the suggestion for d dimensions is to use $X_{1:d} \sim \ iid\ N(0, \sigma^2)$, with $\sigma = 1/\sqrt{d-1}$.

The following code estimates the volume of the unit ball, using iid normal random variables as the sampling distribution, with common mean zero and common standard deviation $\sigma = 1/\sqrt{d-1}$, for $d = 2, 4, 8, 16,$ and 32. Compare this code to the earlier code, which was based on the uniform sampling distribution. Only a few lines are different.

Monte Carlo estimation of the volume of B_d
iid Gaussian sampling distribution

```
% The problem is to estimate the volume of the d-dimensional ball, with
% radius 1 (aka the "unit ball in d dimensions"). The approach is
% importance sampling, which can be viewed as a generalization of rejection
% sampling.

% This example uses the uniform sampling distribution, on (-1,1)^d

% dimensions    A vector--each component of which is a number, d, of
%               dimensions in which we will estimate the volume of the unit
%               ball. Each d assumed to be even, for convenience.

% nsams         The number of samples drawn from the sampling distribution,
%               which is the unifrom distribution on the d-dimensional
%               cube, with sides x_k in (-1,1), for k=1:d.

% nruns         Number of Monte Carlo runs (each using n samples)

dimensions=[2 4 8 16 32];
nsamps=1000000;             % 1,000,000
nruns=3;

ndims=numel(dimensions);
IHat=zeros(nruns,ndims); % Each row is a run and each column is
                               % a number, d,  of dimensions. The entries
                               % are estimated volumes.
CIR=zeros(nruns,ndims);  % Will contain the radiuses of the approximate
                         % 95% confidence intervals, using normal approx.
npos=zeros(nruns,ndims); % How many of the nsampls samples landed in the
                         % unit ball

% Compute the true volumes, using the exact formula for even number of
% dimensions.

HalfD=dimensions/2;
TrueVols=(pi.^HalfD)./factorial(HalfD);

for dim=1:ndims
    d=dimensions(dim);
    for run=1:nruns
        % Sample from the sampling distribution
        sigma=sqrt(1/(d-1)); % mode of radius will be one
        sigma2=sigma^2;
        z=sigma*randn(n,d);  % iid N(0,sigma2)
        % Compute h, which is just the indicator function on the unit ball
        h=sum(z.^2,2)<1;   % h is an nx1 (row) vector, with ones
                           % indicating samples that fall in the unit ball,
                           % and zero indicating those outside ball
        % Compute f
        f=(1/sqrt(2*pi*sigma2))^d * exp(-(sum(z.^2,2)/(2*sigma2))); % the
                % density of the normal dist. in d dimensions, var. sigma2
        % Estimator is mean of h/f
        IHat(run,dim)=mean(h./f);
        % Compute the radiuses of the (approx.) 95% CI's
        SHat=sqrt(mean( (h./f - IHat(run,dim)).^2 )); % Estimated st. dev.
        CIR(run,dim)=2*SHat/sqrt(nsamps); % 2*sigma/sqrt(n)
        % A good measure of efficiency is number of the nsamps that landed
        % in the unit ball
        npos(run,dim)=sum(h);
    end
end
```

The earlier experiment was repeated—three runs, each with sample size $n = 10^6$, for each of the five dimensions. The results were excellent:

d	2	4	8	16	32
$Vol(B_d)$	**3.1416**	**4.9348**	**4.0587**	**0.2353**	**4.3031×10^{-6}**
run 1	3.1346 ± 0.0079	4.9371 ± 0.0123	4.0668 ± 0.0115	0.2353 ± 0.0008	$(4.3003 \pm 0.0177)\times10^{-6}$
run 2	3.1371 ± 0.0079	4.9321 ± 0.0123	4.0580 ± 0.0115	0.2354 ± 0.0008	$(4.3040 \pm 0.0177)\times10^{-6}$
run 3	3.1360 ± 0.0079	4.9303 ± 0.0123	4.0621 ± 0.0115	0.2354 ± 0.0008	$(4.3264 \pm 0.0178)\times10^{-6}$

Furthermore, true to the design of the sampling distribution, the "hit rate" stayed high in the higher dimensions:

d	2	4	8	16	32
run 1	392465	442217	463351	475421	482373
run 2	392939	441935	463289	475474	483378
run 3	392694	441703	463330	474432	484078

2.4 The Bootstrap

Data is expensive. The Bootstrap is a way to make the data work harder. (Resource: Wasserman, Chapter 8.)

Suppose we have five observations from some unknown distribution, F: $x_1 = -1.2$, $x_2 = 7.8$, $x_3 = 2.4$, $x_4 = 3.0$, and $x_5 = -2.1$, where $X_{1:5} \sim iid\ F$ [5]. Consider the following two tasks:

(i) Use the samples to estimate the variance of $X \sim F$. Let's call it σ_F^2. (And, while we're at it, let's write μ_F for the—unknown—mean.)

(ii) Use the samples to estimate the variance of the estimated variance of $X \sim F$. (What does THAT mean?)

Concerning (i), we compute as follows:

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 1.98$$

$$\hat{\sigma}^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^2 = 12.3696$$

Before moving on to (ii), consider this last expression for $\hat{\sigma}^2$: $\hat{\sigma}^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^2$. Could we have written, instead,

$$\hat{\sigma}^2 = \frac{1}{5}\sum_{i=1}^{5} x_i^2 - \left(\frac{1}{5}\sum_{i=1}^{5} x_i\right)^2 ? \tag{5}$$

We have already noted that, in general, $\text{Var}(X) = E[X^2] - E[X]^2$, and this certainly looks a lot like (5).

[5] Recall that $X_{1:5}$ is our shorthand for $(X_1, X_2, X_3, X_4, X_5)$, and that, generically, x_k represents an observation of the random variable X_k.

Of course we could find out by just doing the algebra, but here's another way to look at it which will help when we get to (ii), and maybe even more as we move on to other topics. The idea is to introduce another distribution function, that we will call the *empirical distribution function*. Given a set of n observations, the empirical distribution function is the discrete distribution that is uniform on the observations—each one has probability $\frac{1}{n}$. We will distinguish the empirical distribution from the original distribution by writing $\hat{F}$ instead of just F. Notice that $\hat{F}$ is always discrete, whether or not F is discrete. In our current example, the pmf of $\hat{F}$ is

$$p(x) = \frac{1}{5}\mathbb{1}_{x\in\{x_1,x_2,x_3,x_4,x_5\}}$$

in other words, $p(x) = 0$ unless $x \in \{-1.2, 7.8, 2.4, 3.0, -2.1\}$, in which case $p(x) = \frac{1}{5}$. Alternatively, we could also specify $\hat{F}$ by the cdf:

$$\hat{F}(x) = \frac{\#\{i : x_i \leq x\}}{5}$$

(Make sure you understand this notation, and why $\hat{F}(x)$ is the cdf that corresponds to the pmf $p(x)$.)

Now here's the reason that I am bothering you with these new definitions and notation: instead of writing $\bar{x}$ for $\frac{1}{5}\sum_{i=1}^5 x_i$ we could have written $\frac{1}{5}\sum_{i=1}^5 x_i$ as $E_{\hat{F}}[X]$. (Here I have used $E_{\hat{F}}$ instead of just E so that we don't get confused about which of the two distributions, F or $\hat{F}$, we're working with.) The point is that $\frac{1}{5}\sum_{i=1}^5 x_i$ is *exactly* the expectation of X with respect to the empirical distribution. In fact, to emphasize this, we might even want to write $\mu_{\hat{F}}$ instead of $\bar{x}$. And you can check that, furthermore, $\hat{\sigma}^2$ is exactly $E_{\hat{F}}[(X - E[X])^2]$ (which we could, therefore, reasonably write as $\sigma^2_{\hat{F}}$). This settles the issue: yes, we can write evaluate $\hat{\sigma}^2$ by $\frac{1}{5}\sum_{i=1}^5 x_i^2 - \bar{x}^2$, since this comes down to asking whether $E_{\hat{F}}[(X - E[X])^2] = E_{\hat{F}}[X^2] - E_{\hat{F}}[X]^2$, which is *always* true, for *any* distribution.

Think of it this way: $\hat{F}$ is an approximation of F, based on the available samples. So we are simply approximating μ_F and σ^2_F by $\mu_{\hat{F}}$ and $\sigma^2_{\hat{F}}$! Makes perfect sense.

Now let's give (ii) a try. We not only want to estimate the variance of F from the data, but also estimate the variance of our estimator. In other words, we want to answer this question: how much variation in our estimator of the variance (i.e. in $\hat{\sigma}^2$) would we see *if* we were to recompute it over and over again, each time using five new iid observations from F? After all, $\hat{\sigma}^2$ is a function of $X_{1:5}$, $\hat{\sigma}^2 = \hat{\sigma}^2(X_1, \ldots, X_5)$, and we can expect it to change with every new sample of the five random variables $X_1, \ldots, X_5$. If we run our make-believe experiment B times, and let S_1, S_2, ..., S_B be the B estimated variances, then the answer to (ii) would be the estimated variance of these B numbers:

$$\frac{1}{B}\sum_{b=1}^{B}\Big(S_b - \frac{1}{B}\sum_{d=1}^{B} S_d\Big)^2 \tag{6}$$

which is the estimated variance of $\hat{\sigma}^2(X_1, \ldots, X_5)$ (i.e. the estimated variance of the estimated variance of X!).

All good, except for the fact that nobody wants to collect data over and over again just to get the variation of estimated quantities—they'd rather use all of this new data, all $n = 5B$ samples, to get a better estimate to begin with. And here is where the bootstrap comes in: why not take five new samples, over and over again, from $\hat{F}$ instead of F, since (1) these samples would be cheap, and (2) $\hat{F}$ is, after all, an approximation of F.

Let's generalize. Suppose we have a statistic $S = S(X_1, X_2, \ldots, X_n)$ that is a function of n iid samples from a distribution F. Examples would include the sample variance, $\hat{\sigma}^2$, the median, the sum of the squares, $\sum_{i=1}^{n} X_k^2$, the maximum, the minimum, and so on. For some number B (the bigger the better) we compute B "bootstrap samples" of size n from $\hat{F}$: $X^*_{1:n}(1)$, $X^*_{1:n}(2)$, ..., $X^*_{1:n}(B)$, where for each b, $X_1^*(b), \ldots, X_n^*(b) \sim \; iid \; \hat{F}$.

The corresponding bootstrap samples of S are $S_1^*, S_2^*, \ldots, S_B^*$, where $S_b^* = S(X_1^*(b), X_2^*(b), \ldots, X_n^*(b))$ (or $S(X^*_{1:n}(b))$ for short).

Now we're all set, because from these samples we can estimate (approximate) anything we want about the distribution of $S = S(X_1, X_2, \ldots, X_n)$, including *its* variance—e.g. with equation (6), replacing S_b by S_b^*.

In brief:

$X_1, \ldots, X_n \; \sim \; iid \; F$ (the data)

$\hat{F} \; \sim \; iid \; U\{X_1, \ldots, X_n\}$ (the *empirical distribution*, discrete)

$X^*_{1:n}(1), X^*_{1:n}(2), \ldots, X^*_{1:n}(B) \; \sim \; iid\hat{F}$ (B "bootstrap samples." Each is sampled *with replacement*.)

$S \; = S(X_{1:n})$ (a *statistics*, i.e. a function of the data , such as the median, sample variance, etc.)

$S_b^* \; = S(X^*_{1:n}(b)), \; b = 1, \ldots, B$ (B bootstrap samples of S)

e.g. Suppose that we have n iid observations, $X_1, \ldots, X_n$, from an unknown distribution, F. Let $\mathcal{O}_1 \leq \mathcal{O}_2 \leq \cdots \leq \mathcal{O}_n$ be the corresponding *order statistics*, i.e. the n observations ordered from smallest to largest. Using the order statistics we compute the median of the data:

$$\text{med}(X_{1:n}) = \begin{cases} \frac{\mathcal{O}_{\frac{n}{2}} + \mathcal{O}_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even} \\ \mathcal{O}_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \end{cases}$$

The median, $\text{med}(X_{1:n})$, is a function of the observations, and hence a random variable. How much should we expect it to change from one sample, $X_{1:n}$, to another? Is the value in hand representative of F or likely to be quite different if evaluated on a new sample? Here are two ways to find out:

Method 1 Cheat. Pretend like data doesn't cost anything and choose B iid vectors, each with n iid samples from F. For each of the B vectors, compute the median, and then compute the sample standard deviation of the B medians. This is a "gold standard," since we are looking directly at the variation in the median as a function of the particular sample. It is cheating since we are performing a potentially expensive experiment over and over again, B times.

Method 2 Bootstrap. Use B bootstrap samples of size n, each drawn from $\hat{F}$, which depends only on the *single* set of n variables, $X_1, \ldots, X_n$.

If we pretend like we know the actual distribution, F, then we can perform an experiment comparing the two approaches. As an example, consider the cdf $F(x) = (1 - \frac{1}{x^3})\mathbb{1}_{x>1}$. Read the code entitled "Bootstrap estimation of std of the median," where both methods are implemented with $n = 500$ and $B = 5,000$.

Bootstrap estimation of std of the median

```
% Evaluate the bootstrap for estimating the standard deviation of the
% sample median of the distribution F(x)=1-x^(-3) for x>1, and F(x)=0
% otherwise.

n=500;  % sample size
B=5000; % number of bootstrap samples

% Method 1: Cheat.

% Samples from F can be generated from the uniform distribution, by using
% F^(-1)(U).

U=rand(B,n);                    % Bxn array of iid uniform(0,1) r.v.'s
X=U.^(-1/3);                    % Inverse CDF method for sampling from F.
medians=median(X,2);            % Bx1 array of medians, one for each sample
                                % of n samples - i.e. one for each row of
                                % X.
stdmedian=std(medians);         % The standard deviation of the B samples
                                % of the median.

% Report result

disp(['Direct (cheating) estimate of standard deviation of the median: ',...
    num2str(stdmedian)])

% Method 2: Bootstrap.

% Sample from F, n times

U=rand(1,n);
X=U.^(-1/3);                    % 1xn array of iid r.v.'s from F

% Produce B sample vectors, each of length n, by sampling the components of
% X with replacement

BSS=X(randi(n,B,n));            % Bxn array. Every row is a boostrap sample
                                % from the empirical distribution, \hat{F}.
                                % (In general, randi(m,B,n) produces a Bxn
                                % array of iid random integers, each
                                % uniformly distributed on {1,2,...,m}.)

% Now proceed as though BSS were a Bxn array of iid samples from F (i.e.
% use \hat{F} in place of F, and proceed as we did when we cheated)

medians=median(BSS,2);
stdmedian=std(medians);

% Report result

disp(['Booststrap estimate of standard deviation of the median: ',...
    num2str(stdmedian)])
```

Six runs of the code produced the following output:

```
Direct (cheating) estimate of standard deviation of the median: 0.018752
Booststrap estimate of standard deviation of the median: 0.02092

Direct (cheating) estimate of standard deviation of the median: 0.018876
Booststrap estimate of standard deviation of the median: 0.016792
```

```
Direct (cheating) estimate of standard deviation of the median: 0.018749
Booststrap estimate of standard deviation of the median: 0.019578

Direct (cheating) estimate of standard deviation of the median: 0.018886
Booststrap estimate of standard deviation of the median: 0.023151

Direct (cheating) estimate of standard deviation of the median: 0.018921
Booststrap estimate of standard deviation of the median: 0.01777

Direct (cheating) estimate of standard deviation of the median: 0.018877
Booststrap estimate of standard deviation of the median: 0.024232
```

A glance at the numbers derived from repeatedly sampling F (i.e. cheating) suggests that the true standard deviation is very close to 0.0189. (We could do a lot more cheating to be more certain, with confidence intervals etc., but I've seen enough to bet on the true value being well within, say, 4% of 0.0189.) The bootstrap, which is based on *one* sample of 500 observations, instead of 5,000 samples, typically gets the standard deviation to within about 30%.

2.4.1 A Closer Look at the Bootstrap Sampling Distribution

Each bootstrap sample consists of n independent and identically distributed samples from $\hat{F}$, the uniform distribution on $x_1, x_2, \ldots, x_n$, which themselves resulted from n independent samples from F (the data). Let's assume that $x_1, x_2, \ldots, x_n$ contains no repeats—thre are n distinct values in the data. (A simple modification of the argument here will apply when the sample itself has repeats.)

A single bootstrap sample is a sequence $X^*_{1:n}$ from $U\{x_1, \ldots, x_n\}$, but the actual order of this sequence is usually of no consequence. If, for example, $S(X_{1:n}) = \bar{X}$, then

$$S(X^*_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} X^*_i$$

$$\frac{1}{n}\sum_{k=1}^{n} M_k x_k$$

where M_k is the (random) number of times that the bootstrap sample, $X^*_{1:n}$, chooses x_k:

$$M^*_k \triangleq \#\{i : X^*_i = x_k\}$$

Similarly, if $S(X_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$ (the sample variance), then

$$S(X^*_{1:n}) = \frac{1}{n}\sum_{k=1}^{n} M_k x_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} M_k x_k\right)^2$$

The random variables $M_1, M_2, \ldots, M_n$ have the *multinomial* distribution. Our particular case is a special case of the following general set up for a multinomial distribution: we have an n-sided die, with probabilities p_k, $k = 1, \ldots, n$ of a single roll showing the $k'th$ side (so $\sum_{k=1:n} p_k = 1$). M_k, $k = 1, \ldots, n$

represents the number of k's that appear in N independent rolls. The well-known pmf is

$$\mathbb{P}(M_1 = m_1, \ldots, M_n = m_n) = \begin{cases} \frac{N!}{m_1! m_2! \cdots m_n!} p_1^{m_1} p_2^{m_2} \cdots p_n^{m_n} & \text{if } \sum_{k=1}^n m_k = N \\ 0 & \text{otherwise} \end{cases}$$

The bootstrap is the special case in which $N = n$ and $p_k = \frac{1}{n}$, $k = 1, \ldots, n$. Notice that the multinomial is just a generalization of the binomial, in which a two-sided coin is replaced by an n-sided, not-necessarily fair, die. Notice also that the formulas for $S(X^*_{1:n})$, in terms of $M_1, \ldots, M_n$, are unchanged whether or not the data $x_1, x_2, \ldots, x_n$ has repeated samples.

2.4.2 Bootstrap Confidence Intervals (optional)

There are many ways to use bootstrap samples to construct approximate confidence intervals. Based on experience and some theory, the current recommendations seem to favor what Wasserman calls the "pivotal confidence interval." We are given a sample $X_{1:n}$ from a distribution F, and a statistic $S = S(X_{1:n})$ that estimates a parameter of interest—call it θ, e.g. the median. We seek an approximate confidence interval for θ with confidence level $C \in (0, 1)$. As usual, we start with a collection of bootstrap samples: For some (preferably large) B, we construct B independent random vectors, $X^*_{1:n}(1), \ldots, X^*_{1:n}(B)$, each consisting of n iid samples from $\hat{F}$. To ease notation, let $\hat{\theta}$ be the empirical estimate of θ, i.e. $\hat{\theta} = S(X_{1:n})$.

We will estimate the confidence interval from the "bootstrap errors"

$$\delta_b \triangleq S(X^*_{1:n}(b)) - \hat{\theta}$$

Define α and β to be the left-tail and right-tail $\frac{1-C}{2}$ cutoffs of the set of bootstrap errors:

$$\alpha = \max \left\{ \delta_b : \frac{\#\{\delta_a : \delta_a \le \delta_b\}}{B} \le \frac{1-C}{2} \right\}$$
$$\beta = \min \left\{ \delta_b : \frac{\#\{\delta_a : \delta_a \ge \delta_b\}}{B} \le \frac{1-C}{2} \right\}$$

The bootstrap confidence interval, with confidence level C, is then $(\hat{\theta} - \beta, \hat{\theta} - \alpha)$. (If this looks odd, it might help to consider that α will typically be negative and β will typically be positive.)

2.5 Hypothesis Testing with Random Permutations

Suppose that we want to compare two populations that may or may not have been drawn from the same distribution. For example, we might have a sample of volunteers for a reaction-time experiment. We choose a random subset of the volunteers and give them a low dose of an amphetamine, such as Adderall, while the remainder of the volunteers are given a low dose of an anxiolytic, such as Ativan or Xanax. One hour after administering the drugs, we test the reaction times of all of the individuals.

Let's agree to make the null hypothesis "no effect," meaning that the two populations are essentially indistinguishable in terms of their reaction times. One can imagine many alternative hypotheses of interest,

the most obvious being that, even at these low doses, the reaction times of the amphetamine group are substantially lower than those of the anxiolytic group. Or, perhaps there is reason to believe that the amphetamines will not necessarily lower reaction times, but instead they tend to increase the variance of reaction times, from subject to subject, and/or anxiolytics decrease the variance.

Never mind—for our purpose let's just imagine that we have some statistic that is compellingly relevant to the alternative hypothesis. For example, if the we let $X_1, \ldots, X_m$ be the reaction times of the amphetamine population, and $Y_1, \ldots, Y_n$ be the reaction times of the anxiolytic population, then the statistic might be, simply, $S = S(X_{1:m}, Y_{1:n}) = \bar{X} - \bar{Y}$, the difference between the means. In this case, the alternative hypothesis would likely be that $E[S]$ is less than zero (even at small doses, amphetamines probably *decrease* reaction times and anxiolytics probably *increase* them), whereas, under the null hypothesis, the difference should be zero.

Let $N = m + n$ be the total number of subjects, and define $Z_{1:N}$ by

$$Z_k = \begin{cases} X_k & \text{if } k \leq m \\ Y_{k-m} & \text{if } k > m \end{cases}$$

And let $\mathcal{T}$ be the *set* of all $n+m$ observed reaction times[6]. The idea of the permutation test is to randomly select a subset of size m from the set of $m + n = N$ reaction times. These m times are treated as though they came from the amphetamine subjects, and the remaining n times are treated as though they came from the anxiolytic subjects. Under the null hypothesis, this new ("surrogate") set of reaction times should be indistinguishable from those of the original subjects. In fact, we can do this over and over again, each time evaluating the statistic S, to estimate the *conditional distribution* of S given the set of reaction times in Z, and given that the null hypothesis is true. If the "real" value of the statistic—i.e. S evaluated on the original data—is atypical relative to this distribution, then, under the null something surprising happened, and in fact we can measure how surprising and get a p-value. The beauty of these kinds of tests is that they make almost no assumptions about the distributions of the populations—e.g. no need to assume that they have Gaussian distributions or, for that matter, that they are large enough for the central limit theorem to be relevant.

Let's work through a simple example with $m = 2$ and $n = 3$, which is of course an absurdly small sample. But it does have the advantage that if we rig the numbers right then we can do everything "by hand." We will use the difference of (sample) means for S and pretend that $X_1 = 0.48$ and $X_2 = 0.53$, whereas $Y_1 = 0.55$, $Y_2 = 0.73$, and $Y_3 = 0.58$. These values are certainly rigged, but nevertheless we can ask how surprised we should be if in fact the null hypothesis were correct. The value of the statistic is $S = \bar{X} - \bar{Y} = -0.1150$. Since the amphetamine population has the two fastest reaction times, there is no need to actually build a distribution of S by randomly selecting pairs of reaction times to serve as the surrogate amphetamine population. We can just do the calculation directly: $\mathbb{P}_o(S \leq -0.1150|\mathcal{T}) = \mathbb{P}_o(\text{amphetamine population has the two smallest reaction times}|\mathcal{T}) = 1/\binom{5}{2} = 0.1$. Here, the subscript 'o' indicates that the probabilities are calculated under the null hypothesis, and the conditioning on $\mathcal{T}$ indicates that this is, actually, a *conditional* test—we are making our calculations conditioned on the particular set $\mathcal{T}$.

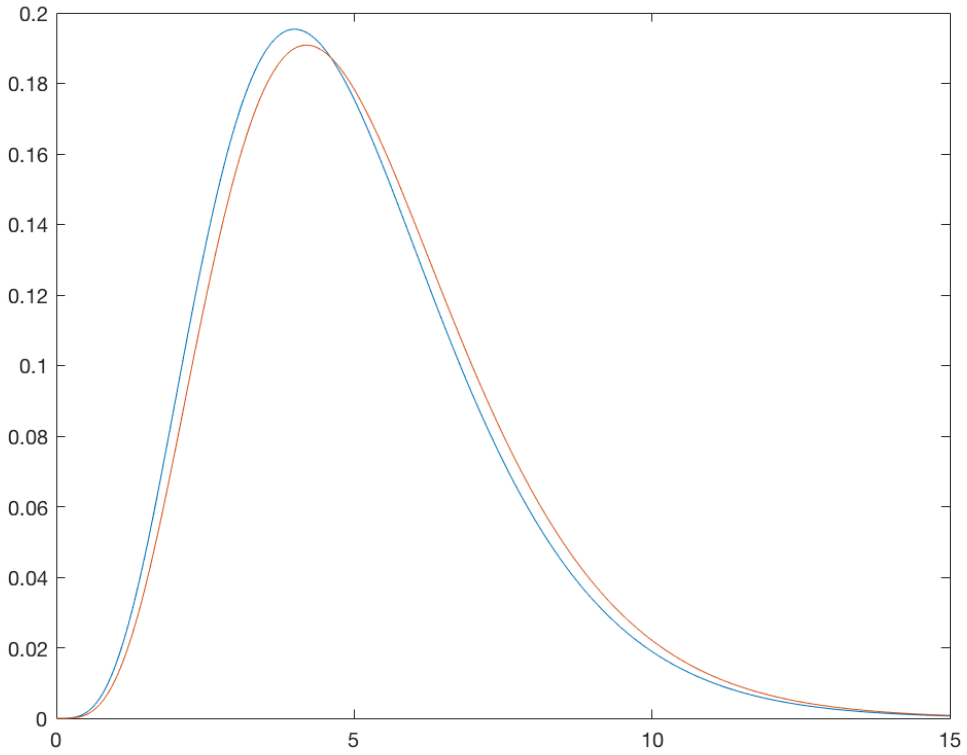
Below you will find code for larger simulations, and its output from a sample run.

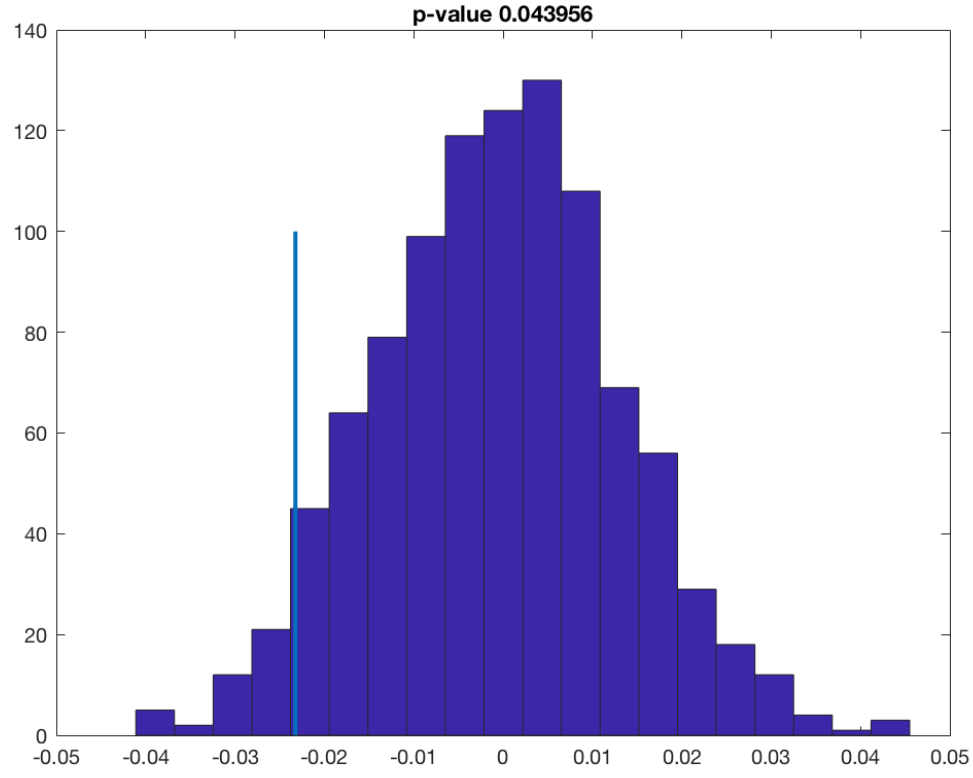
[6]Technically, we want to keep ties, in which case $\mathcal{T}$ might not be a set, since sets can not have repeated elements. The right mathematical structure would be what is known as a *multiset*, with elements and their multiplicities. We will ignore the distinction, though keep in mind that $\mathcal{T}$ may have repeated values.

```
% Let's make some artificial reaction-time data. I will assume that reaction
% times of both groups (amphetamine and anxiolytic) are well modeled by
% scaled gamma distributions.
m=80;               % size of amphetamine population
n=170;              % size of anxiolytic population
B=1000;             % not "B for bootstrap," but it's a similar idea. Here,
                    % B will stand for the number of permutations (= number
                    % of surrogate observations)
% Create make-believe reaction time data, for each of the two populations.
% gamrnd (k,theta,a,b) produces an axb array of iid gamma random variables,
% with common mean k*theta, and common variance k*theta^2
x=gamrnd(5,.5,1,m)/10;          % divided by 10 to put into ballpark of
y=gamrnd(5.7,.5,1,n)/10;        % reaction times, in milliseconds.
% load x and y into a single vector z by concatenation
z=zeros(1,n+m);
z(1:m)=x;
z(m+1:m+n)=y;
% S0 will be the observed value of the statistic (difference in means):
S0=mean(x)-mean(y);
S=zeros(1,B);      % will store the B values of the statistic, one for each
                   % surrogate data set
for per=1:B
    pi=randperm(n+m); % pi(1:(n+m)) is a random permutation of 1,2,...n+m
    xs=z(pi(1:m));    % after permuting, the first m z's go to xs
    ys=z(pi(m+1:m+n));  % and the next n z's go to ys
    S(per)=mean(xs)-mean(ys); % fill S with the values of the statistic for
                              % shuffled groups of amphetamines &
                              % anxiolytics
end
% compute the pvalue
numsmaller=(S<=S0);
pvalue=(sum(numsmaller)+1)/(B+1);
% in figure 1, plot the two densities from which x and y were drawn
v=0:.01:15;
figure(1);
plot(v,gampdf(v,5,1));
hold on
plot(v,gampdf(v,5.2,1));
hold off
% in figure 2, plot the histogram of (surrogate) S values, and indicate
% where S0 (statistic for the observed data) lies
figure(2);
hist(S,20)
v=[S0,S0];w=[0,100];
hold on
plot(v,w,'linewidth',2)
title(['p-value ',num2str(pvalue)])
hold off
```

2.6 The Central Limit Theorem

The central limit theorem allows us to determine the approximate shape of the distribution of our estimators. This gives us a very precise understanding of their accuracy: not only does it give us the mean and variance, but it also tells us something about the magnitude of the "tails." The tails include those very rare events which can be so vitally important in financial markets and other settings.


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15


p-value 0.043956
140
120
100
80
60
40
20
0
-0.05
-0.04
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04
0.05

2.6.1 Refresher: Gaussian Random Variables and Random Vectors

We write $X \sim N(\mu, \sigma^2)$ to indicate that X is a normal (aka Gaussain) random variable with mean μ and variance σ^2. This specification completely defines the distribution of X, which is continuous with pdf

$$f_X(x) = f_X(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The family of Gaussian random variables is tightly knit: If $X \sim N(\mu, \sigma^2)$ and if $Y = aX + b$ then Y is also Gaussian, $Y \sim N(\mu_Y, \sigma_Y^2)$, and we need only determine its mean and variance:

$\mu_Y = E[aX + b] = aE[X] + b = a\mu + b$
$\sigma_Y^2 = \text{Var}[aX + b] = a^2\sigma^2$ (make sure you know why...)

These observations generalize, immediately, to linear combinations of *independent* Gaussian random variables. In fact, if $X_1, \ldots, X_n$ are independent Gaussian, with respective means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, and if $Y = \sum_{i=1}^n a_i X_i + b$, then $Y \sim N(\mu_Y, \sigma_Y^2)$ with

$$\mu_Y = E\big[\sum_{i=1}^n a_i X_i + b\big] = \sum_{i=1}^n a_i \mu_i + b$$
$$\sigma_Y^2 = \text{Var}\big[\sum_{i=1}^n a_i X_i + b\big] = \sum_{i=1}^n a_i^2 \sigma_i^2 \quad \text{(rules of the road!)}$$

(We can go even further: Call $\vec{X} = X_{1:n}$ *jointly Gaussian*, $\vec{X} \sim N(\vec{\mu}, C)$, if

$$f_{\vec{X}}(\vec{x}) = \frac{\exp\Big(-\frac{1}{2}(\vec{x} - \vec{\mu})^T C^{-1} (\vec{x} - \vec{\mu})\Big)}{\sqrt{(2\pi)^n \det C}}$$

where all vectors are $n \times 1$ column vectors, $\vec{\mu} = (\mu_1, \ldots, \mu_n)^T$, C is the $n \times n$ matrix of covariances—$C = \{\text{Cov}[X_i, X_j]\}_{i,j \in \{1,\ldots,n\}}$ and $\text{Cov}[X_i, X_j] = E[(X_i - \mu_i)(X_j - \mu_j)]$ $\forall\, 1 \le i, j \le n$, and $\det C$ is the determinant of C.[7]

In this case, if $\vec{Y} = A\vec{X} + \vec{b}$ for some $m \times n$ matrix A, then $\vec{Y} \sim N(\vec{\mu}_{\vec{Y}}, C_{\vec{Y}})$ where

$\mu_{\vec{Y}} = A\vec{\mu} + b$ (an $m-$dimensional vector)
$C_{\vec{Y}} = ACA^T$ (an $m \times m$ covariance matrix)

Closely knit, indeed.)

[7] We are assuming that $X_1, \ldots, X_n$ are *linearly independent*, i.e. there are not constants $\alpha_1, \ldots, \alpha_n$ and β such that $\sum_{i=1}^n \alpha_i X_i = \beta$ with probability one. Otherwise, some modifications are needed in the formula for $f_{\vec{X}}(\vec{x})$.

2.6.2 Intuition, Formal Statement, Examples

Let $(X(1), \cdots, X(N), \cdots)$ be an infinite sequence of i.i.d random variables, with $E[X(k)] = \mu$ and $\text{var}(X(k)) = \sigma^2$ for every k. Let us consider several ways we can sum these variables together, and what happens when we do.

Averaging Process Let $S_N = \frac{1}{N}\sum_{k=1}^{N} X(k)$ for every natural number, N. Then, as we showed in the previous section, and can be readily reduced from the definition of the expected value and the variance, together with the independence of the variables $X(1) \cdots X(N)$,

$$\begin{aligned} E[S_N] &= \mu && \text{constant} \\ \text{var}(S_N) &= \frac{\sigma^2}{N} && \text{shrinking} \end{aligned}$$

Random Walk Let $S_N = \sum_{k=1}^{N} X(k)$. That is, consider each $X(k)$ as a "step," and let the value S_N be equal to the sum of the first N "steps." Then

$$\begin{aligned} E[S_N] &= N\mu && \text{drifting} \\ \text{var}(S_N) &= N\sigma^2 && \text{expanding} \end{aligned}$$

Centered Random Walk Let $S_N = \sum_{k=1}^{N} (X(k) - \mu)$. In this case, we insist that the average result of each step is zero by subtracting the mean. Then

$$\begin{aligned} E[S_N] &= 0 && \text{constant} \\ \text{var}(S_N) &= N\sigma^2 && \text{expanding} \end{aligned}$$

In the three examples above, we see that the variance of the sum is either shrinking to zero or expanding to infinity. Is there something in-between? Is there a process of this form such that the expected value and variance are simply constant for all N? The answer is yes.

Central Limit Theorem Process Let $S_N = \frac{1}{\sqrt{N}}\sum_{k=1}^{N} (X(k) - \mu)$. Then

$$\begin{aligned} E[S_N] &= 0 && \text{constant} \\ \text{var}(S_N) &= \sigma^2 && \text{constant} \end{aligned}$$

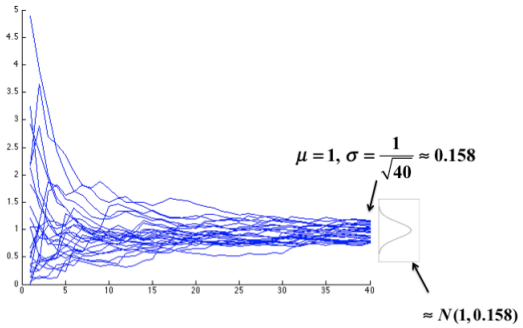
In this case the distribution of S_N converges as $N \to \infty$ to something quite specific: $\mathcal{N}(0, \sigma^2)$.

Theorem. *(Central Limit Theorem). Let $X(1), X(2) \cdots$ be an infinite sequence of i.i.d random variables with mean μ and variance σ^2. Let $S_N = \frac{1}{\sqrt{N}}\sum_{k=1}^{N} (X(k) - E[X(k)])$. Then the distribution of S_N is asymptotically $\mathcal{N}(0, \sigma^2)$. In particular, if $Z \sim \mathcal{N}(0, \sigma^2)$, then*

$$\begin{aligned} \lim_{N\to\infty} \mathbb{P}(a < S_N < b) &= \mathbb{P}(a < Z < b) \\ &= \int_a^b \frac{e^{-\frac{z^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dz \end{aligned}$$

The central limit theorem is a statement about the central limit theorem process, but it can help us think about the random walk process and the averaging process.

Let $X(1), X(2) \cdots$ i.i.d $\sim f(x) = e^{-x}$. Then $E[X(k)] = \text{var}(X(k)) = 1$. Let S_N denote an averaging process for these variables, $S_N = \frac{1}{N}\sum_{k=1}^{N} X(k)$. The law of large numbers tells us that $S_N \to E[X(1)] = 1$. Elementary calculations will show that $\text{var}(S_N) = 1/N$. The central limit theorem tells us that the approximate shape of the distribution of the variable S_N is Gaussian.



Monte-Carlo estimators are based on the law of large numbers, so the central limit theorem also helps us determine the reliability of Monte-Carlo estimators. Let $\hat{I}_N$ denote an unbiased estimator of the form $\hat{I}_N = \frac{1}{N}\sum_{k=1}^{N} Y(k)$, where $Y(k)$ are i.i.d, $E[Y(k)] = I$, and $\text{var}(Y(k)) = \sigma^2$. Let Z be a gaussian random variable, $\mathcal{N}(I, \sigma^2/N)$. According to the central limit theorem, $\hat{I}_N$ and Z have roughly the same distribution. Therefore,

$$\begin{aligned}\mathbb{P}(I - \epsilon < \hat{I}_N < I + \epsilon) &\approx \mathbb{P}(I - \epsilon < Z < I + \epsilon) \\ &= \int_{I-\epsilon}^{I+\epsilon} \frac{e^{-\frac{(z-I)^2}{2\sigma^2/N}}}{\sqrt{2\pi\sigma^2/N}} dz\end{aligned}$$

This 1-dimensional integral can be computed easily. These probabilities give us approximate rates of convergence for Monte-Carlo estimators.

In order for this theorem to work, the mean and variance must be well-defined. This is not always the case. For example, the Cauchy distribution is given by the probability density function

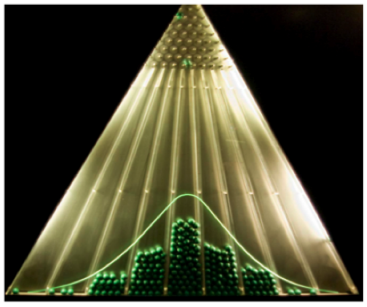
$$f_X(x) = \frac{1}{\pi + \pi x^2}$$

This is a perfectly valid distribution. However, if $X \sim f_X$ then

$$\begin{aligned}E[X] &= \int f_X(x) \cdot x \cdot dx = \int_{-\infty}^{\infty} \frac{x}{\pi + \pi x^2} dx \\ &= \int_{-\infty}^{0} \frac{x}{\pi + \pi x^2} dx + \int_{0}^{\infty} \frac{x}{\pi + \pi x^2} dx \\ &= -\infty + \infty = \text{undefined}\end{aligned}$$

Thus the expected value of X is undefined. Therefore the variance (which is defined in terms of the mean) is also undefined. Neither the central limit theorem nor the law of large numbers apply to this

v. The "bean machine" demonstrates the CLT



Each peg moves ball right or left by one unit, with equal probabilities.

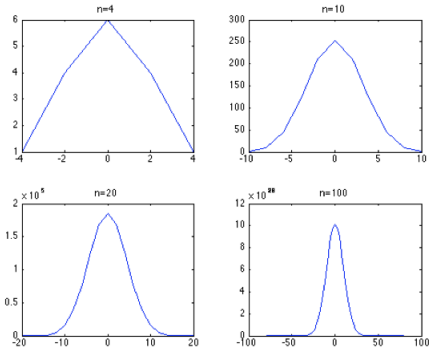
Mimicking repeated samples of $\sum_{k=1}^{n} X_k$, where $\mathbb{P}(X_k = -1) = \mathbb{P}(X_k = +1) = 0.5$.

Figure 1: The Bean Machine

case. In fact, if $X(1), X(2) \cdots$ is an infinite sequence of random variables with this distribution, and $S_N = \frac{1}{N} \sum_{k=1}^{N} X(k)$, then S_N will never converge.

Why does the central limit theorem work? Basically, it's about *combinatorics*. There are more paths to the values that are closer to the mean, and the number of paths to a given value outweighs the details of the probability of X.

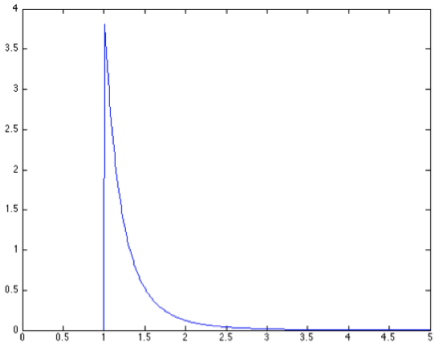
In the bean machine, assume that there are N rows of pegs, so that each ball hits N pegs and moves either one unit to the right or one unit to the left. If a particular ball goes right in k out of the N collisions, then its final position, relative to the center, is at $2k - N$, and there are $\binom{N}{k}$ ways to get there. For each of $N \in \{4, 10, 20, 100\}$, here are plots of $\binom{N}{k}$ as a function of $2k - N$ where $k = 0, 1 \cdots N$.



When $N = 100$, the shape of this distribution is indistinguishable from a Gaussian curve.

The *rate* of convergence to the normal distribution can depend on the distribution of X. For example, if we choose $X(1) \cdots X(N)$ to be normally distributed, then $S_n = \sum X_k$ is always normally distributed, for every N. However, sampling from highly skewed distributions or distributions with "heavy tails" can make for slow convergence. These observations are especially true if we measure the rate of convergence by *ratios* of probabilities.

Example Let $X(1) \cdots X(N)$ i.i.d $\sim$ Pareto(C, α). This distribution carries the density $f(x) = \mathbb{1}_{x>C} \alpha C^\alpha / x^{\alpha+1}$. For example, taking $C = 1$ and $\alpha = 4$, we obtain



This distribution has var$(X(1)) = 2/9$ and $E[X(1)] = 4/3$. Let

$$\bar{X} = \frac{1}{100} \sum_{k=1}^{100} X(k)$$

According to the central limit theorem, we would expect that $\bar{X}$ would be roughly distributed like $\mathcal{N}(4/3, 2/900)$. Thus, $\left(\bar{X} - 4/3\right) / \sqrt{2/900}$ should be distributed like $\mathcal{N}(0, 1)$. For example, if $Z \sim \mathcal{N}(0, 1)$, then

$$\begin{aligned}
\mathbb{P}(\bar{X} < 1) &= \mathbb{P}\left(\frac{\bar{X} - 4/3}{\sqrt{2/900}} < \frac{1 - 4/3}{\sqrt{2/900}}\right) \\
&\overset{CLT}{\approx} \mathbb{P}\left(Z < \frac{1 - 4/3}{\sqrt{2/900}}\right) \\
&= \mathbb{P}(Z < 7.0711) \\
&= 7.69 \times 10^{-13}
\end{aligned}$$

This number[8] is clearly wrong. Since $X(k) \geq 1$ for all k, we know that $\mathbb{P}(\bar{X} < 1) = 0$. Similarly, for the right tail:

$$\begin{aligned}
\mathbb{P}(\bar{X} > 1.4) &= \mathbb{P}\left(\frac{\bar{X} - 4/3}{\sqrt{2/900}} > \frac{1.4 - 4/3}{\sqrt{2/900}}\right) \\
&\overset{CLT}{\approx} \mathbb{P}\left(Z > \frac{1.4 - 4/3}{\sqrt{2/900}}\right) \\
&= 1 - \mathbb{P}\left(Z < \frac{1.4 - 4/3}{\sqrt{2/900}}\right) \\
&= .0786
\end{aligned}$$

This number is also wrong. We conducted a simple experiment, in which we calculated 10^6 samples of $\bar{X}$ (this involves drawing 100×10^6 samples from the Pareto distribution). From these samples, we obtained a

[8] The numerical value of $\mathbb{P}(Z < 7.0711)$ can be readily obtained in any programming language. For example, in Matlab, the command `normcdf`(-7.0711) will yield the correct result.

95% confidence interval for $\mathbb{P}(\bar{X} > 1.4) = E\left[\mathbb{1}_{\bar{X}>1.4}\right]$, given by $(.0832, .0843)$. This suggests the central limit theorem underestimated the right tail of the distribution by roughly 6%.

2.6.3 Convolutions

When $X(1), X(2) \cdots$ are independent and identically distributed, the central limit theorem gives the approximate distribution of $\sum_{k=1}^{N} X(k)$ – but what is the true distribution?

The sum of two independent random variables is calculated through a technique called convolution. For example, suppose $X \sim f_X$ and $Y \sim f_Y$ are independent. Then the cumulative distribution function of $S = X + Y$ may be calculated by integration:

$$\begin{aligned} F_X(s) &\triangleq \mathbb{P}(S < s) \\ &= \int \int_{\{(x,y):\ x+y<s\}} f_X(x) f_Y(y) dy dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) \int_{y=-\infty}^{s-x} f_Y(y) dy dx \end{aligned}$$

The density of S may be calculated by deriving the cumulative distribution function. Since all the functions are well-behaved, we can take this derivative inside the integral:

$$\begin{aligned} f_S(s) &= \int_{x=-\infty}^{\infty} f_X(x) \frac{d}{ds} \left(\int_{y=-\infty}^{s-x} f_Y(y) dy \right) dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) f_Y(s-x) dx \end{aligned}$$

The last expression is known as the "convolution" of f_X and f_Y.

Definition. *Let f_X and f_Y denote any two functions. Then the **convolution** of f_X and f_Y is denoted by $h = (f_X * f_Y)$ and is defined by*

$$h(s) = (f_X * f_Y)(s) = \int_{x=-\infty}^{\infty} f_X(x) f_Y(s-x) dx$$

Convolutions carry a few nice properties:

- Commutativity – $f_X * f_Y = f_Y * f_X$
- Associativity – $(f_X * f_Y) * f_Z = f_Y * (f_X * f_Z)$
- Differentiability – the convolution $f_X * f_Y$ is differentiable if either f_X or f_Y is differentiable. If both are differentiable, then

$$\left(\left(\frac{d}{dx} f_X \right) * f_Y \right)(s) = \left(f_X * \left(\frac{d}{dy} f_Y \right) \right)(s) = \frac{d}{ds} (f_X * f_Y)(s)$$

We can use convolution iteratively to determine the distribution of $\sum_{k=1}^{N} X(k)$. Let $X(k)$ i.i.d $\sim f$, and let g_N denote the density of $S_N = \sum_{k=1}^{N} X(k)$. Then $S_1 = X(1)$, so $g_1(s) = f(s)$. Since $S_2 = X_2 + S_1$,

$$g_2(s) = (f * g_1)(s)$$

Since $S_3 = X_3 + S_2$

$$gf(s) = (f * g_2)(s)$$

and so-on. In general

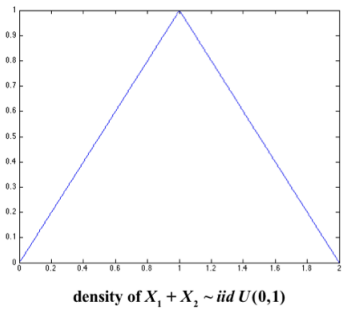
$$g_N(s) = \left(\overbrace{f * f * f * f * \cdots f}^{\text{n times}} \right)$$

gives the density of S_N.

Example Let $f(x) = \mathbb{1}_{x\in[0,1]}$, the density of the uniform distribution $U[0,1]$. Then

$$\begin{aligned} g_2(s) &= \int_{-\infty}^{\infty} \mathbb{1}_{x\in[0,1]} \mathbb{1}_{s-x\in[0,1]} dx \\ &= \int_0^1 \mathbb{1}_{s-x\in(0,1)} dx = \int_0^1 \mathbb{1}_{x\in(s-1,s)} dx \\ &= \begin{cases} 0 & \text{if } s \le 0 \\ s & \text{if } 0 < s \le 1 \\ 2-s & \text{if } 1 < s \le 2 \\ 0 & \text{if } s > 2 \end{cases} \end{aligned}$$

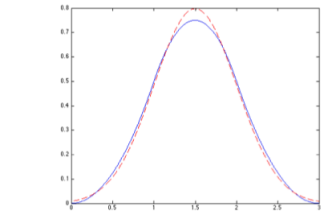
That is, g_2 is the triangle density:



Continuing,

$$\begin{aligned} g_3(s) &= \int_{-\infty}^{\infty} \mathbb{1}_{x\in[0,1]} f_2(s-x) \\ &= \cdots = \begin{cases} 0 & \text{if } s \le 0 \\ s^2/2 & \text{if } 0 < s \le 1 \\ 1 - \frac{2-s^2}{2} - \frac{1-s^2}{2} & \text{if } 1 < s \le 2 \\ 1 - \frac{3-s^2}{2} & \text{if } 2 < s \le 3 \\ 0 & \text{if } s > 3 \end{cases} \end{aligned}$$

Which is piecewise quadratic:



density of $X_1 + X_2 + X_3 \sim iid\ U(0,1)$ (blue, solid)
density of normal with mean $\mu = 1.5$ & variance $\sigma^2 = 0.25$ (red, broken)

This is already quite Gaussian in shape. It is tedious to continue integrating by hand. However, a computer makes it quite easy, as you will observe in the homework.

We conclude this section with a caution. When $X \sim \text{Uniform}[0, 1]$, we have seen that the central limit theorem converges extremely rapidly. For more skewed or heavy-tailed distributions, this convergence can become arbitrarily slow. Unfortunately, many Monte-Carlo estimators are both skewed and heavy-tailed!

3 Estimation

- Our focus will be on *nonparameteric* (as opposed to *parametric*) estimation
- We will learn some cool tricks (linearization and the support-vector machines, cross validation, etc.)
- There will be some cautionary tales (the bias-variance "dilemma," the "curse of dimensionality")
- We will develop methods that find there way into many modern applications, including: image and language "parsing"; image processing (de-noising, in-painting, etc.); image analysis (face detection, license-plate reading, etc.); neuroscience (man-machine interfaces, deciphering the "neural code," etc.); "quant trading" (e.g. statistical arbitrage); computational biology (deciphering gene-regulatory networks, gene-disease associations, protein folding, etc.).

Let $X_1, \ldots, X_n \sim \ iid\ f(x;\theta)$ where θ is an unknown parameter. the goal is to estimate θ using $X_{1:n}$.

$\theta \in \mathbb{R}^d$ ($d < \infty$) "parametric estimation"

$\theta \in \mathbb{R}^\infty$ "non-parametric estimation"

3.1 Bias and Consistency

Generically, we will write $\hat{\theta} = \hat{\theta}_n = \hat{\theta}_n(X_{1:n})$ for an estimator of θ based on $X_{1:n}$. When talking about estimators, we usually talk in terms of some basic properties. We call $E[\hat{\theta}_n] - \theta$ the *bias'*, and when $\hat{\theta} = \theta$ we say that $\hat{\theta}$ is *unbiased*. If $lim_{n\to\infty}\hat{\theta}_n = \theta$ then we say that "$\hat{\theta}_n$ is *consistent*".

The last concept, that of consistency, is a little bit subtle. What exactly do we mean by $\hat{\theta}_n \to \theta$? Here are three (of many) possibilities:

$$\mathbb{P}(\hat{\theta}_n \to \theta) = 1 \text{ "almost sure," or "with prob. one"} \tag{7}$$

$$\forall \epsilon > 0\ \mathbb{P}(|\hat{\theta}_n - \theta| < \epsilon) \xrightarrow{n\to\infty} 1 \text{ "in probability"} \tag{8}$$

$$E[|\hat{\theta}_n - \theta|^2] \xrightarrow{n\to\infty} 0 \text{ "in mean square"} \tag{9}$$

Almost sure (7) is generally considered the strongest, since it implies convergence in probability (8) and, when $E[\hat{\theta}_n^2] < \infty$, it also implies mean square convergence (9). We won't make much use of almost sure convergence since it involves measure theory and, besides, the distinction between the three is not generally of any practical importance. But both probability and mean square convergence are intuitive, and in fact closely related. We've already used mean square convergence in one version of of the LLN: $X_{1:n} \sim iid$ with common mean μ and common variance $\sigma^2 < \infty \Rightarrow$

$$\begin{aligned} E\Big[|\frac{1}{n}\sum_{k=1}^n X_k - \mu|^2\Big] &= \text{Var}\Big(\frac{1}{n}\sum_{k=1}^n X_k - \mu\Big) \\ &= \text{Var}\Big(\frac{1}{n}\sum_{k=1}^n (X_k - \mu)\Big) \\ &= \frac{\sigma^2}{n} \to 0 \end{aligned}$$

In other words, $\frac{1}{n}\sum_{k=1}^{n} X_k \to \mu$ in mean square (9). It turns out that (9) $\Rightarrow$ (8), in which case we get another form of the LLN:

$$\forall \epsilon > 0 \ \mathbb{P}(|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu| < \epsilon) \to 1$$

The implication (9) $\Rightarrow$ (8) is the consequence of a handy relationship known as the Markov Inequality:

Proposition. *Let W be a non-negative random variable ($\mathbb{P}(W \geq 0) = 1$). Then for any $\epsilon > 0$, $\mathbb{P}(W \geq \epsilon) \leq \frac{E[W]}{\epsilon}$.*

This is easy to prove. Let's do the special case, W continuous with $W \sim f$: fix $\epsilon > 0$

$$\begin{aligned}
\mathbb{P}(W \geq \epsilon) &= \int_{x=\epsilon}^{\infty} f(x)dx \\
&\leq \int_{x=\epsilon}^{\infty} \frac{x}{\epsilon} f(x)dx \\
&= \frac{1}{\epsilon}\int_{x=\epsilon}^{\infty} x f(x)dx \\
&\leq \frac{1}{\epsilon}\int_{x=0}^{\infty} x f(x)dx = \frac{E[W]}{\epsilon}
\end{aligned}$$

Now assume $E[|\hat{\theta}_n - \theta|^2] \xrightarrow{n\to\infty} 0$ (convergence in mean square). Then

$$\begin{aligned}
1 - \mathbb{P}(|\hat{\theta}_n - \theta| < \epsilon) &= \mathbb{P}(|\hat{\theta}_n - \theta| \geq \epsilon) \\
&= \mathbb{P}(|\hat{\theta}_n - \theta|^2 \geq \epsilon^2) \\
&\leq \text{(Markov Inequality)} \frac{E\left[|\hat{\theta}_n - \theta|^2\right]}{\epsilon^2} \to 0
\end{aligned}$$

In other words, (9) $\Rightarrow$ (8).

e.g. $X \sim N(\mu, 1)$, $\hat{\mu}_n(X_{1:n}) = \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$. Check that $\hat{\mu}_n$ is unbiased and consistent for μ.

e.g. $X \sim N(\mu, \sigma^2)$

$\hat{\mu}_n(X_{1:n}) = \bar{X}$

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$$

Verify that both $\hat{\mu}_n$ and $\hat{S}_n^2$ are unbiased and consistent, for μ and σ^2 respectively.

Notice that we have replaced the usual $\frac{1}{n}$ by $\frac{1}{n-1}$ in the definition of $\hat{S}_n^2$. This doesn't effect consistency, since the ratio, $\frac{n-1}{n} \to 1$. But it is necessary if we want an unbiased estimator of σ^2. One way to arrive at this is to compute the expected value of the "$\frac{1}{n}$" estimator ($\hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2$), and then correct its

bias. Start with a simple identity (see equation (5), and the discussion of the empirical distribution):

$$E\big[\frac{1}{n}\sum_{k=1}^{n}(X_k-\bar{X})^2\big] = E\big[\frac{1}{n}\sum_{k=1}^{n}X_k^2 - \big(\frac{1}{n}\sum_{k=1}^{n}X_k\big)^2\big]$$
$$= E\big[\frac{1}{n}\sum_{k=1}^{n}X_k^2\big] - E\big[\big(\frac{1}{n}\sum_{k=1}^{n}X_k\big)^2\big]$$

Let's take these one at a time:

$$E\big[\frac{1}{n}\sum_{k=1}^{n}X_k^2\big] = \frac{1}{n}\sum_{k=1}^{n}E[X_k^2] = E[X^2]$$

where $X \sim N(\mu,\sigma^2)$. As for the second term:

$$\begin{aligned}
E\big[\big(\frac{1}{n}\sum_{k=1}^{n}X_k\big)^2\big] &= E\big[\big(\frac{1}{n}\sum_{k=1}^{n}X_k\big)\big(\frac{1}{n}\sum_{l=1}^{n}X_l\big)\big] \\
&= E\big[\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}X_kX_l\big] \\
&= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}E[X_kX_l] \\
&= \frac{1}{n^2}\sum_{k=1}^{n}E[X_k^2] + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l\neq k,\ l=1}^{n}E[X_kX_l] \\
&= \frac{1}{n}E[X^2] + \frac{n(n-1)}{n^2}E[X_1]E[X_2] \\
&= \frac{1}{n}E[X^2] + \frac{n-1}{n}E[X]^2
\end{aligned}$$

And now, putting everything together:

$$\begin{aligned}
E[\hat{\sigma}_n^2] &= E\big[\frac{1}{n}\sum_{k=1}^{n}(X_k-\bar{X})^2\big] \\
&= E\big[\frac{1}{n}\sum_{k=1}^{n}X_k^2\big] - E\big[\big(\frac{1}{n}\sum_{k=1}^{n}X_k\big)^2\big] \\
&= E[X^2] - \frac{1}{n}E[X^2] - \frac{n-1}{n}E[X]^2 \\
&= \frac{n-1}{n}\big(E[X^2]-E[X]^2\big) = \frac{n-1}{n}\sigma^2
\end{aligned}$$

In other words $\hat{\sigma}_n^2$ is biased; it is just a little bit too small, on average. But this is easy to fix: Let $\hat{S}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k-\bar{X})^2$. Then

$$E[\hat{S}_n^2] = \frac{n}{n-1}E\big[\frac{1}{n}\sum_{k=1}^{n}(X_k-\bar{X})^2\big] = \frac{n}{n-1}E[\hat{\sigma}_n^2] = \big(\frac{n}{n-1}\big)\big(\frac{n-1}{n}\big)\sigma^2 = \sigma^2$$

There is another way to look at this that will help you to understand why $\hat{\sigma}_n^2$ was doomed to be, on average, too small. Consider that, among all numbers α, $\alpha = \bar{X}$ is the number that minimizes $\sum_{k=1}^n (X_k - \alpha)^2$, as you can easily verify using a little bit of calculus. Consequently, unless $\bar{X}$ happens to equal μ,

$$\begin{aligned}
\sigma^2 &= E[(X-\mu)^2] \\
&= E[\frac{1}{n}\sum_{k=1}^n (X_k - \mu)^2] \\
&> E[\frac{1}{n}\inf_\alpha \sum_{k=1}^n (X_k - \alpha)^2] \\
&= E[\frac{1}{n}\sum_{k=1}^n (X_k - \bar{X})^2]
\end{aligned}$$

The change $\frac{1}{n} \to \frac{1}{n-1}$ makes $\frac{1}{n}\sum_{k=1}^n (X_k - \bar{X})^2$ bigger by just the right amount.

As a final remark, concerning the last two examples, none of our calculations used the assumption that we had sampled from a Gaussian distribution. The statements about bias and consistency hold for *any* distribution with finite variance ($\sigma^2 < \infty$).

Let's look now at a non-parametric example:

e.g. At a different extreme, we might make no assumptions at all about the distribution of the data: $X_{1:n} \sim \overline{iid}\ F$ for some unknown cdf F. We don't even assume that F is discrete or continuous—it might be neither. To connect to the previous notation, we might write $F(x) = F(x;\theta)$, but now θ is everything—the entire distribution, $\theta = F$!

We have already worked with a natural estimator for F: the uniform distribution on the samples, $X_1, \ldots, X_n$, which we called the "empirical distribution," and denoted it by $\hat{F}_n$.

If we are going to be dealing with estimators that are themselves distributions (rather than just a finite number of parameters), then we might be better off introducing the probability, on $\mathbb{R}$, that is "induced" by X: $P = P_X$, where $P(B) = \mathbb{P}(X \in B)$, for (almost) all $B \subseteq \mathbb{R}$. The empirical distribution function, F is often called, simply, the distribution of X, and the same goes for P—it is often referred to as the distribution of X. What is important is that P is indeed a legitimate probability distribution on a particular Ω, namely $\Omega = \mathbb{R}$.

Unfortunately, all of the above variations appear in one place or another within the textbooks and literature, and you will have to learn to be flexible.

The analog of the empirical distribution, $\hat{F}_n$, is the empirical probability $\hat{P}_n$, which refers to the same distribution, the one that is uniform on the samples. $\hat{P}_n$ is sometimes written like this: $\hat{P}_n = \frac{1}{n}\sum_{k=1}^n \delta_{X_k}$, where δ_{x_o}, the "delta function" at x_o, represents the trivial distribution that is x_o with probability 1. In other words, $\delta_{x_o}(B)$ is one if $x_o \in B$ and zero otherwise. Keeping this in mind, the formula for $\hat{P}_n$ should now make sense: for any set $B \subseteq \mathbb{R}$,

$$\hat{P}_n(B) = \frac{1}{n}\sum_{k=1}^n \delta_{X_k}(B) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{X_k \in B}$$

which is just the number of samples, $X_1, \ldots, X_n$, that are in B, divided by n. In other words, we have a new way to write the same distribution that we formerly denoted as $\hat{F}_n$.

It's kind of like electronic medical records. A good idea, but then everyone decided to use their own format. (The unfortunate result is that no data is harder to wrangle than medical data.)

How good is $\hat{P}_n$? For starters, it is unbiased, at least in the following sense: for (almost) any $B \subseteq \mathbb{R}$

$$\begin{aligned} E\left[\hat{P}_n(B)\right] &= E\left[\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{X_k \in B}\right] \\ &= \frac{1}{n}\sum_{k=1}^{n} E\left[\mathbb{1}_{X_k \in B}\right] \\ &= \frac{1}{n}\sum_{k=1}^{n} \mathbb{P}(X_k \in B) \\ &= \mathbb{P}(X_k \in B) = P(B) \end{aligned}$$

So where are we? Well, for one thing we are in infinite dimensions, trying to estimate an entire probability distribution. Luckily, we have a natural estimator, the empirical distribution $\hat{P}_n$ (aka $\hat{F}_n$), and it is unbiased for (almost) any set B in $\mathbb{R}$: $E[\hat{P}_n(B)] = P(B)$. How about consistency? Following this notion of unbiased, we might hope for success with a similar notion of consistency: does $\hat{P}_n(B) \to P(B)$ for each B? Yes, by yet another application of the law of large numbers:

$$\begin{aligned} \hat{P}_n(B) &= \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{X_k \in B} \\ &\overset{n\to\infty}{\to} E\left[\mathbb{1}_{X \in B}\right] \text{ (LLN)} \\ &= \mathbb{P}(X_k \in B) = P(B) \end{aligned}$$

Remarks (optional)

i. It is tempting to get greedy, and ask that the convergence of $\hat{P}_n(B)$ to $P(B)$ be *uniform* in B. For example, we might hope that

$$\sup_B \left|\hat{P}_n(B) - P(B)\right| \overset{n\to\infty}{\longrightarrow} 0 \text{ in probability}$$

But this would indeed be going to far. For a simple counterexample, suppose that F is continuous. Now, given $X_{1:n}$, consider the particular set $\tilde{B} = \{X_1, \cdots, X_n\}$. Since $\tilde{B}$ has only a finite number of points, $P(\tilde{B}) = 0$. At the same time, however, $\hat{P}_n$ is the empirical distribution and it puts *all* of its mass on $\{X_1, \cdots, X_n\}$, so $\hat{P}_n(\tilde{B}) = 1$. Consequently $\sup_B \left|\hat{P}_n(B) - P(B)\right| >= \left|\hat{P}_n(\tilde{B}) - P(\tilde{B})\right|$ is always one, and certainly not going to zero.

ii. Still, it is at least a little surprising that there is an infinitely large set of B's on which $\hat{P}_n(B)$ *does* converge to $P(B)$ uniformly: Let $\mathcal{B} = \{B : B = (-\infty, x], \text{ for some } x \in \mathbb{R}\}$. Then

$$\sup_{B\in\mathcal{B}} \left|\hat{P}_n(B) - P(B)\right| \overset{n\to\infty}{\longrightarrow} 0 \text{ in probability}$$

This is known as the Glivenko-Cantelli theorem, and may be written in terms of $\hat{F}_n$ as

$$\sup_{x\in\mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \overset{n\to\infty}{\longrightarrow} 0 \text{ in probability}$$

3.2 Performance Measures and the Bias/Variance Tradeoff

There are many possible measures of performance, and we can expect the best choice to be problem dependent. But in terms of mathematical convenience and insight, a very good place to start is with the time-honored notion of mean squared error (MSE).

3.2.1 MSE for Parametric Models

Let's return to where we started with our discussion of estimation: Let $X_1, \ldots, X_n \sim \ iid\ f(x;\theta)$ where θ is unknown and the object of interest. Assume, further, that $\theta \in \mathbb{R}^1$, and that $\hat{\theta}$ (=$\hat{\theta}_n(X_{1:n})$) is an estimator of θ.

Definition.

$$\textit{MSE}_\theta(\hat{\theta}) \triangleq E_\theta \left[\left(\hat{\theta} - \theta \right)^2 \right]$$

The mean squared error can be broken down into the sum of two very informative pieces:

$$\begin{aligned}
MSE_\theta(\hat{\theta}) &= E_\theta \left[\left(\left(\hat{\theta} - E_\theta \left[\hat{\theta} \right] \right) + \left(E_\theta \left[\hat{\theta} \right] - \theta \right) \right)^2 \right] \\
&= E_\theta \left[\left(\hat{\theta} - E_\theta \left[\hat{\theta} \right] \right)^2 + \left(E_\theta \left[\hat{\theta} \right] - \theta \right)^2 + 2 \left(E_\theta \left[\hat{\theta} \right] - \theta \right) \left(\hat{\theta} - E_\theta \left[\hat{\theta} \right] \right) \right] \\
&= \text{Var}[\hat{\theta}] + \left(E_\theta \left[\hat{\theta} \right] - \theta \right)^2 + 2 \left(E_\theta \left[\hat{\theta} \right] - \theta \right) E_\theta \left[\left(\hat{\theta} - E_\theta \left[\hat{\theta} \right] \right) \right] \\
&= \text{Var}[\hat{\theta}] + \left(\underbrace{E_\theta \left[\hat{\theta} \right] - \theta}_{\text{“bias”}} \right)^2
\end{aligned}$$

In summary, MSE=Var+Bias^2. This is the "bias-variance tradeoff." A good estimator must have both small variance and small bias, but as we will see shortly, and after that again and again, these two conditions tend to work against each other. In any case, the message for now, is that if we want mean-square consistency, then we will need to have both the bias and the variance going to zero as $n \to \infty$.

e.g. Let's go back to the simple example $X_1 \cdots X_n \sim \mathcal{N}(\mu, \sigma^2)$. That is, let $X_1 \cdots X_n$ be independent and identically distributed according to density $f(x;\mu,\sigma^2) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$, where and μ, σ are unknown parameters. Now let's take a close look at the performance of the various estimators we have for means and variances:

1. $\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k$. We have already noted that $\hat{\mu}$ is *unbiased* (bias is zero): $E[\hat{\mu}] = \mu$. Since the variance is given by $\text{Var}[\hat{\mu}] = \sigma^2/n$, the MSE is exactly the same thing as the variance.

2. $\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} (X_k - \hat{\mu})^2$. We've already calculated the mean of $\hat{\sigma}^2$:

$$E\left[\hat{\sigma}^2\right] = \frac{n-1}{n}\sigma^2$$

   So $\hat{\sigma}^2$ is biased. In particular, it has a negative bias:

$$\begin{aligned} E\left[\hat{\sigma}^2\right] - \sigma^2 &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &= -\frac{1}{n}\sigma^2 \end{aligned}$$

3. $\hat{S}^2 = \frac{1}{n-1}\sum_{k=1}^{n} (X_k - \hat{\mu})^2$, which turned out to have no bias: $E\left[\hat{S}^2\right] = \sigma^2$.

It turns out that the variance of both $\hat{\sigma}^2$ and $\hat{S}^2$ go to zero as $n \to \infty$. And since both are asymptotically unbiased, they are both mean square (aka MSE) consistent:

$$E\left[(\hat{\sigma}^2 - \sigma^2)^2\right] \overset{n\to\infty}{\longrightarrow} 0 \quad \text{and} \quad E\left[(\hat{S}^2 - \sigma^2)^2\right] \overset{n\to\infty}{\longrightarrow} 0$$

So which is better? Possibly, your main interest would be in the MSE, not so much asymptotically (infinity is pretty far away), but at finite n. Here's where we can make very good use of the assumption $X \sim \mathcal{N}(\mu, \sigma^2)$ In general, a "which is better" question like this one will have a "well, it depends" type answer—namely, it depends on the true underlying distribution. We can, however, get a complete answer in the particular case of sampling from a Gaussian distribution.

The computations behind getting the variance and mean squared errors of these two estimators are quite tedious and overwhelmingly error prone, even in the Gaussian case, so let's just jump to a summary:

	$E\left[(\hat{\sigma}^2 - \sigma^2)^2\right]$	$E\left[(\hat{S}^2 - \sigma^2)^2\right]$
bias^2	$\left(-\frac{\sigma^2}{n}\right)^2$	0
variance	$\left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1}$	$\frac{2\sigma^4}{n-1}$
MSE	$\frac{2n-1}{n^2}\sigma^4$	$\frac{2}{n-1}\sigma^4$

With a little bit of algebra, you should be able to convince yourself that $\frac{2n-1}{n^2}\sigma^4 < \frac{2}{n-1}\sigma^4$, and hence $\hat{\sigma}^2$ has the better MSE performance. Evidently, in going from $\hat{S}^2$ to $\hat{\sigma}^2$ we make a good trade: the squared bias goes up a bit, but the variance goes down by more.

Remarks Here are a few other things that are known about the problem of estimating the variance of a Gaussian distribution:

1. $\hat{S}^2$ has the lowest variance among all *unbiased* estimators. (Notice that this does not contradict the conclusion above, since $\hat{\sigma}^2$ is biased.)

2. Among all estimators of the form $b\hat{S}^2$, for some constant b, the (MSE) best is $\frac{n-1}{n+1}\hat{S}^2$. Neither of the two that we analyzed!

3.2.2 Mean Squared Error for some Non-parametric Models

What, if anything, is the natural extension of the concept of mean squared error to the nonparametric case? There is no one answer, but let's look at two examples and, for each, define a natural performance measure based on squared differences between estimators and the quantities being estimated.

e.g. 1 Consider a discrete distribution on the positive integers: $f_X(k) = p_k$, for all $k \in \mathbb{N} \triangleq \{1, 2, \ldots\}$. We wish to estimate f_X, which amounts to estimating the sequence $p_1, p_2, \ldots$ under the constraint that $p_k \geq 0,\ \forall k \in \mathbb{N}$ and $\sum_{k=1}^{\infty} p_k = 1$.

Given $X_{1:n} \sim\ iid\ f_X$, let $\hat{p}_k(X_{1:n})$, for each $k \in \mathbb{N}$, be the estimated value of p_k. The squared error for $\hat{p}_k$ is simply $(\hat{p}_k - p_k)^2$, but this doesn't really capture the global problem, which is to estimate an *infinite* number of parameters $(p_1, p_2, \ldots)$, not just one. What about replacing the squared error with the *sum* of squared errors ("SSE"):

$$\text{SSE} = \text{SSE}(X_{1:n}) = \sum_{k=1}^{\infty} (\hat{p}_k - p_k)^2$$

With this notion of error, our performance measure could then be the mean (expected value) of the SSE (the "MSSE"):

$$\begin{aligned} \text{MSSE} &= E\big[\sum_{k=1}^{\infty} (\hat{p}_k - p_k)^2\big] \\ &= \sum_{k=1}^{\infty} E\big[(\hat{p}_k - p_k)^2\big] \end{aligned}$$

by linearity. The summands, $E\big[(\hat{p}_k - p_k)^2\big]$, $k \in \mathbb{N}$, are mean squared errors, in the parametric sense, and hence admit the bias-variance decomposition:

$$E\big[(\hat{p}_k - p_k)^2\big] = \big(E[\hat{p}_k(X_{1:n})] - p_k\big)^2 + \text{Var}[\hat{p}_k(X_{1:n})]$$

Here, I have shown explicitly the dependence on the data $X_{1:n}$ to emphasize that the expected value and variance are taken with respect to these random variables—i.e. with respect to the true (and unknown) distribution f_X. Putting this decomposition back into the formula for the MSSE gives us a global bias-

variance type of decomposition:

$$\begin{aligned}\text{MSSE} &= \sum_{k=1}^{\infty} E\big[(\hat{p}_k - p_k)^2\big] \\ &= \sum_{k=1}^{\infty} \big(E[\hat{p}_k(X_{1:n})] - p_k\big)^2 + \sum_{k=1}^{\infty} \text{Var}[\hat{p}_k(X_{1:n})] \\ &= \sum_{k=1}^{\infty} \text{bias}(k)^2 + \sum_{k=1}^{\infty} \text{var}(k)\end{aligned}$$

Let's see what this looks like for a particular estimator: $\hat{p}_k(X_{1:n}) = \frac{1}{n}\sum_{l=1}^{n} \mathbb{1}_{X_l=k}$, which is just the relative frequency estimator, $\hat{p_k} = N_k/n$, where N_k counts the number of k's in the sample. Notice that N_k is a binomial random variable, with parameters n and p_k (flip a biased coin, with probability p_k of heads, n times and count the number, N_k, of heads). Recall (or convince yourself) that $N \sim$ binomial(n, p_k) has mean np_k and variance $np_k(1-p_k)$. Hence

$$\big(E[\hat{p}_k(X_{1:n})] - p_k\big)^2 = \big(E[\frac{N_k}{n}] - p_k\big)^2$$
$$= \big(\frac{np_k}{n} - p_k\big)^2 = 0$$

and

$$\text{Var}[\hat{p}_k(X_{1:n})] = \text{Var}[\frac{N_k}{n}] = \frac{p_k(1-p_k)}{n} \quad \text{(why?)}$$

Imagine that, we have an exact expression for our measure of performance in a non-parametric setting (don't get used to it):

$$\text{MSSE} = \frac{1}{n}\sum_{k=1}^{\infty} p_k(1-p_k) = \frac{1}{n}\sum_{k=1}^{\infty} p_k - \frac{1}{n}\sum_{k=1}^{\infty} p_k^2 = \frac{1}{n}\big(1 - \sum_{k=1}^{\infty} p_k^2\big)$$

Among other things, we can see that the MSEE is bounded above by $\frac{1}{n}$, independent of the particular distribution, $p_1, p_2, \ldots$ on $\mathbb{N}$. The best case (lowest upper bound) occurs when $p_k = 1$ for some k, since then MSEE will be zero, right out of the gate. What's the worst case? (Hint: there is no "worst case," but for any fixed n we can find a pmf on $\mathbb{N}$ that puts the MSSE arbitrarily close to $\frac{1}{n}$. How?)

e.g. 2 Often we are willing to assume that the data is drawn from a continuous distribution, in which case there is a pdf, $f_X(x)$, which we may or may not want to model parametrically. Let's assume that there is a lot of data and we would prefer to make as few *a priori* assumptions as possible.

In the next section we will look at a well-known approach to non-parametric density estimation, but for now let's jump ahead and assume that we have a sample $X_1, X_2, \ldots, X_n \sim \ iid\ f$ and an estimator $\hat{f} = \hat{f}_{X_{1:n}}$. We want to identify a useful measure of performance. Following the previous example, it would be natural to begin by looking at the squared difference between $\hat{f}$ and f at a single $x \in \mathbb{R}$: $\big(\hat{f}(x) - f(x)\big)^2$. As in the discrete case, we would like a global measure, rather than one that depends on a single value of x. But here, instead of using the sum over k we use the integral over x:

$$ISE \triangleq \int \Big(\hat{f}(x) - f(x)\Big)^2 dx$$

the "integrated squared error." The *mean* integrated squared error is then

$$E_f\left[\int\left(\hat{f}(x)-f(x)\right)^2 dx\right]$$

and it has a bias-variance decomposition based on the point-wise decomposition of $E\left[\left(\hat{f}(x)-f(x)\right)^2\right]$, as follows:

$$\begin{aligned}
MISE &= E\left[\int_x\left(\hat{f}(x)-f(x)\right)^2 dx\right] \\
&= \int_x E\left[\left(\hat{f}(x)-f(x)\right)^2\right] dx \\
&= \int_x\left(E\left[\hat{f}(x)\right]-f(x)\right)^2 dx + \int_x \text{Var}[\hat{f}(x)]dx \\
&= \int \text{bias}(x)^2 dx + \int \text{var}(x)dx
\end{aligned}$$

where $\text{bias}(x) = E\left[\hat{f}(x)\right]-f(x)$ and $\text{var}(x) = \text{Var}\left[\hat{f}(x)\right]$. As usual, we have suppressed the dependency of $\hat{f}$ on the data $X_{1:n}$ (replacing $\hat{f}(x;X_{1:n})$ with $\hat{f}(x)$), but we must keep in mind that expectations and variances are taken with respect to $X_{1:n}$, and under the true distribution f.

Next, we will explore the kernel density estimator, a particular approach to non-parametric density estimation. We will demonstrate a stark trade-off between bias and variance, and, from this point of view, explore the popular method of "data-driven smoothing" known as cross-validation.

3.3 Kernel Density Estimation

3.3.1 Elements of a Kernel Estimator

Given $X_1, X_2, \ldots, X_n \sim\ iid\ f$, where f is a probability density function, assumed to be unknow. We want to construct an estimator, $\hat{f}$ ($=\hat{f}(x;X_{1:n})$ which is MISE consistent for f.

We will build $\hat{f}$ out of scaled and shifted copies of a *kernel* κ(x), where

1. $$\kappa(x)\geq 0,\ \int\kappa(x)=1,\ \int\kappa(x)x=0\ \int\kappa(x)x^2=1$$

   In other words, κ is itself a density function with mean zero and variance 1. And example would be

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

   the pdf of a standard normal ($\mathcal{N}(0,1)$).

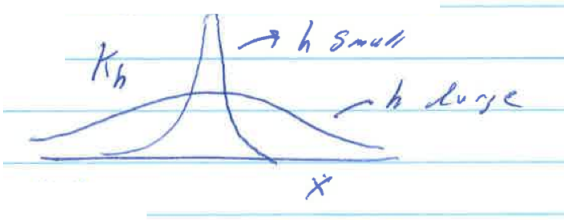
2. Let $\kappa_h = \kappa(x/h)/h$. Convince yourself that for any x_o, $\kappa_h(x - x_o)$ is again a pdf, this time with mean x_o and standard variance h^2 (standard deviation h).

3. Let $\hat{f}_h$ denote the kernel density estimator defined by

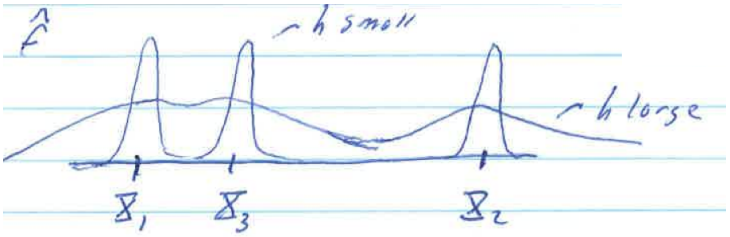
$$\hat{f}_h(x) = \frac{1}{n}\sum_{k=1}^{n} \kappa_h(x - X_k)$$

a sum of copies of κ centered at the data, scaled by h, and weighted by $\frac{1}{n}$.

The value of h controls the "bandwidth" or "smoothness" of the kernel:



Depending on h, one can obtain very different estimates for f:



3.3.2 Bias versus Variance

Let us see how h alters the bias and variance of our estimator.

Bias As $h \to 0$, we obtain the empirical distribution. Intuitively speaking, this estimator has "zero bias."

$$\begin{aligned}\lim_{h\to 0}\left(E\left[\hat{f}_h(x)\right] - f(x)\right) &= \lim_{h\to 0}\int f(t)\kappa_h(x-t)dt - f(x) \\ &= \lim_{h\to 0}\left(\kappa_h * f\right)(x) - f(x) = 0 \end{aligned} \tag{10}$$

On the other hand, as $h \to \infty$,

$$\lim_{h\to\infty}\left(E\left[\hat{f}_h(x)\right] - f(x)\right) = \lim_{h\to\infty}\left(\kappa_h * f\right)(x) - f(x) = 0 - f(x) \tag{11}$$

Important Aside: Notice that

$$E\left[\hat{f}_h(x)\right] = \left(\kappa_h * f\right)(x)$$

and recall that the right hand side, which is the convolution of f with κ_h, represents the density of the *sum of two independent random variables, one drawn from f and the other from κ_h*. The bias, at x, comes from the *blurring* of the density of X at x, and this blurring can be though of as the result of adding noise to X, where the noise has density κ_h. When h is small, the noise is typically small and the blurred density at x will not be much different from $f(x)$. But if h is large, then the variance of the noise is large and the blurring will be quite substantial. Keep this in mind—it's a good way to look at the source of bias in kernel estimation.

As usual, our interest is not so much in the bias at a single x, but in the integrated squared bias. We can make use of the two calculations above, equations (10) and (11), to conclude:

$$\int \text{bias}\left(\hat{f}(x)\right)^2 dx \to \begin{cases} 0 & \text{as } h \to 0 \\ \int f(x)^2 dx & \text{as } h \to \infty \end{cases}$$

(The argument here is not quite rigorous, but it is close enough, and in any case it is true with minimal assumptions.)

Variance By contrast, the variance vanishes for large values of h. Indeed, as $h \to \infty$, the kernel density estimator does not depend on the data at all, and so the variance is very small indeed! Conversely, as h gets smaller and smaller, the estimator becomes highly dependent on the exact positions of the observations—so much so that the variance diverges to infinity as $h \downarrow 0$.

For those interested, here are the details, but this material is decidedly optional:

Derivation of the variance of the kernel estimator (optional) In general, for any fixed value of x,

$$\begin{aligned}
\text{Var}[\hat{f}_h(x)] &= \text{Var}\left[\frac{1}{n}\sum_{k=1}^{n} \kappa_h(x - X_k)\right] \\
&= \frac{1}{n}\text{Var}\left[\kappa_h(x - X)\right] \\
&= \frac{1}{n}\int f(t)\left(\kappa_h(x-t)\right)^2 dt - \frac{1}{n}\left(\int f(t)\kappa_h(x-t)dt\right)^2 \\
&= \frac{1}{n}\underbrace{\int f(t)\left(\kappa_h(x-t)\right)^2 dt}_{A_h(x)} - \frac{1}{n}\left(\underbrace{\left(f * \kappa_h\right)(x)}_{B_h(x)}\right)^2
\end{aligned}$$

And so the overall variance is bounded by

$$\frac{1}{n}\int A(x)dx - \frac{1}{n}\int B(x)^2 dx \leq \int \text{Var}[\hat{f}_h(x)]dx \leq \frac{1}{n}\int A(x)dx$$

Let us consider the integral for each term, $A_h(x)$ and $B_h(x)^2$.

1. The integral of $A_h(x)$ can be computed by changing the order of integration:

$$\begin{aligned}
\int A_h(x)dx &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)\left(\kappa_h(x-t)\right)^2 dtdx \\
&= \int_{-\infty}^{\infty} f(t)\int_{-\infty}^{\infty} \kappa_h(x-t)^2 dxdt \\
&= \int_{-\infty}^{\infty} f(t)\int_{-\infty}^{\infty} \kappa_h(x)^2 dxdt \\
&= \int_{-\infty}^{\infty} \kappa_h(x)^2 dx \\
&= \frac{1}{h^2}\int_{-\infty}^{\infty} \kappa\left(\frac{x}{h}\right)\times\kappa\left(\frac{x}{h}\right)dx \\
&= \frac{1}{h}\int_{-\infty}^{\infty} \left(\kappa(x)\right)^2 dx
\end{aligned}$$

2. The integral of $B_h(x)$ can be bounded by an inequality due to Young, one form of which is this: for any densities f and κ_h,

$$\int \left((f*\kappa_h)(x)\right)^2 dx \le \int \left(f(x)\right)^2 dx$$

Putting the pieces together, we have that

$$\frac{\frac{1}{h}\int \kappa(x)^2dx - \int f(x)^2dx}{n} \le \int \text{Var}[\hat{f}_h(x)]dx \le \frac{1}{nh}\int \kappa(x)^2dx$$

**********end optional**********

The result is that

$$\int \text{Var}\left[\hat{f}(x)\right]dx \to \begin{cases}\infty & \text{as } h\to 0\\ 0 & \text{as } h\to\infty\end{cases}$$

Let's look at the behavior of the bias and variance next to each other. For any fixed number of samples, n, and any kernel, κ, we have shown that

	$h\to 0$	$h\to\infty$
$\int \text{bias}(\hat{f}(x))^2dx$	$\to 0$	$\to \int f(x)^2dx$
$\int \text{Var}[\hat{f}(x)]dx$	$\to\infty$	$\to 0$

This does not look so good. We can choose between zero variance with substantial bias or zero bias with infinite variance!

3.4 Data-driven Smoothing

Sometimes we refer to the bandwidth, h, as the "smoothing parameter." This makes sense: as we increase h the estimator gets smoother and smoother. (And don't forget that the variance gets smaller and smaller and the bias gets bigger and bigger.) Data-driven smoothing is a generic term for a collection of approaches that have one thing in common: they all use the data itself to try and determine a good value for the smoothing parameter.

But before we get to any examples, the first thing to know is that there are "good values" for h. By this I mean that if, instead of giving up because of the depressing looking table above, we do some mathematics, then things might not look so bleak. For example, what if we ask an oracle (e.g. maybe your favorite deity) for the very best h for any given data set $X_{1:n}$. Since the oracle knows all, and in particular knows f, she (my favorite happens to be female) might give you the value of h^* that minimizes the MISE:

$$h^* \triangleq \operatorname*{argmin}_h E\Big[\int_{x=-\infty}^{\infty} \Big(\hat{f}_h(x) - f(x)\Big)^2\Big]$$

(argmin_h means the value of h that minimizes the expression that follows, in other words, just what we want). Of course it is important to remember that h^* will depend on $X_{1:n}$ *and* f (that's where the oracle comes in), which we could emphasize by writing $h^*(X_{1:n}, f)$ instead of just h^*.

The point of all this is that we can prove, under mild conditions on f and κ, that

(i) $E\Big[\int_{x=-\infty}^{\infty} \Big(\hat{f}_{h^*(X_{1:n},f)}(x) - f(x)\Big)^2\Big] = \mathcal{O}(n^{-4/5}) \to 0$ (the oracle's choice gives us a consistent nonparametric density estimator, and we even know how fast it converges to the true density)

(ii) $h^*(X_{1:n}, f) = \mathcal{O}(n^{-1/5})$ (we also know how fast the best h goes to zero)

In fact, even without knowing f, if we take $h = h_n$ and let it go to zero slowly enough (e.g. $h_n = \frac{c}{n^{1/5}}$ for any constant c) then $\hat{f}_{h_n} \to f$ in the MISE sense.

Not that any of this solves our *practical* problem, which is to choose a good value for the smoothing parameter given a fixed, finite, sample of data. But it does tell us that the situation is not hopeless and it might even encourage us to look for ways to choose h from the data—"data-driven smoothing". Here are three:

3.4.1 Cross-validated Likelihood

This is one of many related techniques that use the basic principle of constructing estimators on part of the data and using the rest of the data to estimate generalization performance and adjust parameters. Cross-validation techniques usually do this many times, often by leaving out random subsets of a given size, and aggregating the results by one means or another.

Cross-validation for bandwidth selection in a kernel density estimator is an excellent example. I will discuss two varieties, one here and one in the next section, and you will have a good opportunity to gain some intuition and insight through computational experiments in Assignment 5.

The idea of cross-validated likelihood (sometimes called "ordinary cross validation") begins with the construction of an estimator from all but one data value: for a fixed $l \in \{1, \ldots, n\}$ pretend as if X_l were lost or somehow deleted from the data set. Let $X_{1:n}^{(l)}$ represent the set of observations excluding the single observation X_l: $X_{1:n}^{(l)} \triangleq \{X_1, \ldots, X_{l-1}, X_{l+1}, \ldots, X_n\}$. Now define $\hat{f}_h^{(l)} = \hat{f}_h^{(l)}(x; X_{1:n}^{(l)})$ to be the kernel estimator based on the incomplete data $X_{1:n}^{(l)}$:

$$\hat{f}_h^{(l)}(x; X_{1:n}^{(l)}) = \frac{1}{n-1} \sum_{k=1, k\neq l}^{n} \kappa_h(x - X_k)$$

By construction, $\hat{f}_h^{(l)}$ really knows nothing about X_l, in fact $\hat{f}_h^{(l)}$ is *independent* of X_l. What should we expect of $\hat{f}_h^{(l)}$ if we were to evaluate it at X_l? Since the data is assumed to be iid, we might as well think of X_l as a future observation, as though we never had it in the first place. Since the whole idea is to perform well in the future, let's try to use X_l to evaluate performance. Here's where the likelihood principle comes in: a density $g(x)$ evaluated at $x = X_l$ is sometimes called the likelihood of X_l under g. The reason for the name is that if $g(X_l)$ is large then the hypothesis that $X_l \sim g$ makes sense—it was anticipated, likely, or at least not unlikely. On the other hand, if $g(X_l)$ is small, then X_l was *unanticipated* in the sense of being *unlikely* if we were to believe that $X_l \sim g$. This idea of evaluating a model (g) by the likelihood of the data ($g(X_l)$) under g is the reasoning behind the estimation method known as maximum likelihood estimation. We will return to maximum likelihood shortly (§3.5), and explore some more mathematical and rigorous justifications.

For now, I hope that you will agree that the size of $\hat{f}_h^{(l)}(X_l; X_{1:n}^{(l)})$ is, at the least, a sensible (albeit weak) measure of how well $\hat{f}_h^{(l)}$ fits the (unknown) true pdf, f. As such, it is also a measure of the effectiveness of the bandwidth, h. Of course we don't want to make too much of the evaluation of an estimated density at a single point, but why not look at $\hat{f}_h^{(l)}(X_l; X_{1:n}^{(l)})$ for *every* $l \in \{1, 2, \ldots, n\}$, since the same reasoning applies at each of the n data points. This leads to a likelihood-like expression (sometimes called a "pseudo-likelihood")

$$\tilde{J}(h) = \prod_{l=1}^{n} \hat{f}_h^{(l)}(X_l; X_{1:n}^{(l)})$$

as a candidate for a data-driven measure of the effectiveness of any given bandwidth h. Ordinary cross validation refers to choosing the bandwidth h by maximizing $\tilde{J}(h)$. Then, putting everything together, we arrive at a fully data-driven estimator $\hat{f}_{\hat{h}}$, where $\hat{h} = \operatorname{argmax}_h \tilde{J}(h)$, and

$$\hat{f}_{\hat{h}}(x) = \frac{1}{n} \sum_{k=1}^{n} \kappa_{\hat{h}}(x - X_k)$$

Usually, for computational simplicity, we work with the logarithm of likelihoods rather than likelihoods themselves. Since the logarithm is a monotone and strictly increasing function, $\operatorname{argmax}_h \tilde{J}(h) = \operatorname{argmax}_h \log \tilde{J}(h)$. And finally, for the purpose of unifying the notation between this section and the next, we set $J(h) = -\log \tilde{J}(h)$ and refer to

$$\hat{h} \triangleq \operatorname*{argmin}_h J(h)$$

as the (ordinary) cross-validated choice of the smoothing parameter h. (Don't let the notational changes bother you; we are still just maximizing $\tilde{J}(h)$.) See Assignment 5 for some experimental results. There

are also theoretical guarantees, but they are quite restrictive. Nevertheless this has not kept practitioners from using the method, often with good success.

3.4.2 Cross-validated ISE, and the Benefits of Regularization

We seek a good value of h for the kernel estimator

$$\hat{f}_h(x) = \hat{f}_h(x; X_{1:n}) = \frac{1}{n}\sum_{k=1}^{n} \kappa_h(x - X_k) \tag{12}$$

In theory, we would like to pick h to minimize the integrated squared error. That is, we would like to take h to be

$$\operatorname*{argmin}_h \text{ISE}(\hat{f}_h) = \operatorname*{argmin}_h \int_{x=-\infty}^{\infty} (f(x) - \hat{f}_h(x))^2 dx \tag{13}$$

Of course we cannot compute the minimizer (13) because we do not know f to begin with—if we did then we wouldn't be using kernel density estimators to estimate it.

Instead, let's break down the integrated squared error into pieces, and see if we can estimate the pieces, one at a time:

$$\int_x (f(x) - \hat{f}_h(x))^2 dx = \int_x \hat{f}_h(x)^2 dx \tag{14}$$

$$-2\int_x \hat{f}_h(x) f(x) dx \tag{15}$$

$$+\int_x f(x)^2 dx \tag{16}$$

The last term, (16), would be a show-stopper, except that we don't care about it at all. We are trying to minimize ISE with respect to h and happily (16) doesn't depend on h. One down, two to go.

The first term, (14), isn't much trouble either. After all, for each value of h we have an explicit formula for $\hat{f}_h(x)$, namely (12). All we have to do is square and integrate. Of course the integral might be unpleasant, or even intractable, but we can always resort to a Riemann approximation (which is exactly what you will do in Assignment 5) or, if we happen to be working in high dimensions we might be best off using Monte Carlo integration (of all things). So let's assume that we've done the work and have (14) as a function of h.

That leaves the middle term, (15), which is problematic. In particular in involves f, the very thing we're trying to estimate. Here there's a nice trick, and it connects the two approaches, ordinary cross validation and cross-validated ISE. It turns out that the very same leave-one-out trick that we used in the previous section is the key to usefully estimating (15). Specifically, we can replace $\int_x \hat{f}_h(x) f(x) dx$ by the random variable (function of $X_1, \ldots, X_n$)

$$\frac{1}{n}\sum_{l=1}^{n} \hat{f}_h^{(l)}(X_l) \tag{17}$$

With a little bit of algebra you can convince yourself that (17) and (15) have exactly the same expected values. This might seem a bit thin as an argument for replacing one with the other, but at the same time

we're a bit desperate. All of which leads us to define:

$$J(h) \triangleq \int_x \hat{f}_h(x)^2 dx - 2\frac{1}{n}\sum_{l=1}^{n} \hat{f}_h^{(l)}(X_l)$$

as an approximation for the h-dependent part of the integrated squared error; $\hat{h} = \text{argmin}_h J(h)$ as our data-driven smoother; and $\hat{f}_{\hat{h}}$ as our estimator for f. (Note that $\frac{1}{n}\sum_{l=1}^{n} \hat{f}_h^{(l)}(X_l)$ can be rewritten as $\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,l\neq k}^{n} \kappa_h(X_l - X_k)$, which is how it appears in Assignment 5.)

But does it make any sense, and does it work? The answer to both is yes. As for whether it makes sense, J has an intuitive interpretation that is independent of its derivation as an approximation of ISE. Let's take the second term first: as discussed previously, it makes sense to choose a value of h that makes the leave-one-out data as likely as possible. The only difference is that here we aggregate the individual likelihoods, $\hat{f}_h^{(l)}(X_l),\ l = 1, \ldots, n$, by averaging rather than multiplying as in ordinary cross validation. Either way, it's a good idea to adjust h to make this term large (and hence minus two times this term small).

Concerning the first term, if it were not for the squared integrand, it would always be one, since $\hat{f}_h$ is a density for every h. By squaring the integrand, the integral becomes a collective measure of the sizes of the peaks of $\hat{f}_h$—small h makes for big peaks, and big peaks make for large values of the integral. Without this term, effectively a "regularization" term (in this case, a so-called L_2 regularization), the minimum of J is achieved at very small values of h. The resulting estimator, $\hat{f}_{\hat{h}}$, performs poorly.

But the properly regularized estimator is effective, as you will see from the assignment. There is, as well, a strong theoretical justification, due to a remarkable result by Charles Stone. Stone's theorem asserts that not only is $\hat{f}_{\hat{h}}$ MISE consistent, but the rate at which the integrated squared error, $\int_x (\hat{f}_{\hat{h}}(x) - f(x))^2 dx$, goes to zero is as fast as if $\hat{h}$ were chosen, instead, to minimize (13). In other words, this version of cross validation works as well, asymptotically, as if were granted the *true* ISE for each h and the privilege to minimize it. It is perhaps a little strange that this kind of estimator is almost unknown in the machine learning community.

3.5 Maximum Likelihood

Let's return to the general problem of parametric estimation: we are given an iid sample, $X_1, \ldots, x_n$, from a pmf or pdf $f = f(x;\theta)$. The parameter, $\theta \in \mathbb{R}^d$ ($\theta = (\theta_1, \ldots, \theta_d)$), is unknown. The goal is to estimate θ from the data, $X_{1:n}$.

Among the many approaches, one stands out as the most general and, in terms of efficiently using the data, the most powerful. This is the maximum likelihood approach. Many of the classical estimators (e.g. the sample mean or sample variance) and many of the newer, big-data-style, estimators can be viewed, one way or another, as maximum likelihood estimators or clever approximations of maximum-likelihood estimators. The basic idea is always the same: Choose θ to make the observed data $X_{1:n}$ as plausible, aka likely, as possible.

3.5.1 The Likelihood Function and the Maximum-Likelihood Estimator

We are assuming that the observed data is drawn iid from $f(\cdot, \theta)$. Therefore, the *joint* density or probability mass function of the data will be a product of the densities or probability mass functions evaluated at the data. Recall, for example, that if f is a pmf (the distribution is discrete) than

$$\mathbb{P}(X_1 = x1, \ldots, X_n = x_n) = \prod_{k=1}^{n} f(x_k; \theta)$$

If the distribution is continuous, then $\prod_{k=1}^{n} f(x_k; \theta)$ is no longer a probability, but it is the next best thing: the joint density of the n random variables, $X_{1:n}$ evaluated at $x_{1:n}$. The main point is that in both cases the representation for the joint probability, whether through a pmf or a pdf, is a product, every term of which depends on the unknown parameter θ.

Definition. *The function*

$$L = L(x_1, \ldots, x_n; \theta) \triangleq= \prod_{k=1}^{n} f(x_k; \theta)$$

is called, variously, the data likelihood*, the* likelihood of the data*, or the* likelihood of the parameters $\theta_1, \ldots, \theta_d$*.*

Maximum likelihood estimation chooses θ to maximize the likelihood function:

Definition. *The following function $\hat{\theta}$, where*

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) \triangleq \operatorname*{argmax}_{\theta} \prod_{k=1}^{n} f(x_k; \theta)$$

is called the maximum-likelihood estimator *(MLE) of θ, and sometimes denoted $\hat{\theta}_{ML}$.*

In fact, it is almost always more convenient to work with the logarithm of the likelihood function, rather than the likelihood itself:

Definition. *The* log likelihood function*, which we will denote $\tilde{L}$, is the function*

$$\tilde{L} = \tilde{L}(x_1, \ldots, x_n; \theta) \triangleq= \sum_{k=1}^{n} \log f(x_k; \theta)$$

Since the logarithm is monotonic, and strictly increasing (for positive arguments), $\operatorname*{argmax}_{\theta} \prod_{k=1}^{n} f(x_k; \theta)$ and $\operatorname*{argmax}_{\theta} \sum_{k=1}^{n} \log f(x_k; \theta)$ are the same thing. We can use whichever one is easier to work with.

3.5.2 Examples

(i) **(The normal distribution)** Consider the density of the normal distribution with mean μ and variance σ^2. The likelihood function, given $X_1 = x_1, \ldots, X_n = x_n \sim \ iid \setminus(\mu, \sigma^2)$ is

$$\begin{aligned} L(x_{1:n}; \theta) &= L(x_{1:n}; \mu, \sigma^2) \\ &= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\sum_k \frac{(x_k-\mu)^2}{2\sigma^2}} \\ &\Rightarrow \\ \tilde{L} &= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{k=1}^{n} \frac{(x_k-\mu)^2}{2\sigma^2} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2 \end{aligned}$$

Set the partial derivatives to zero:

$$\underline{\frac{\partial \tilde{L}}{\partial \mu} = 0}$$

$$0 = \frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k-\mu) \Rightarrow \sum_{k=1}^{n}(x_k-\mu) = 0 \Rightarrow \hat{\mu}_{ML} = \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{x}$$

As for σ^2, let's let $\gamma = \sigma^2$ to avoid confusion about the derivatives:

$$\underline{\frac{\partial \tilde{L}}{\partial \gamma} = 0}$$

$$0 = \frac{\partial}{\partial \gamma}\left(\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\gamma) - \frac{1}{2\gamma}\sum_{k=1}^{n}(x_k-\bar{x})^2\right) = -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{k=1}^{n}(x_k-\bar{x})^2$$

$$\Rightarrow \hat{\gamma}_{ML} = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2 \Rightarrow \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2$$

(ii) **(The exponential distribution)** Let $f(x;\alpha) = \alpha e^{-\alpha x}\mathbb{1}_{x>0}$. Estimate α given a sample $X_1 = x_1, \ldots, X_n = x_n$. Notice that it is safe to assume that all of the samples are positive, since $\mathbb{P}(X \leq 0) = 0$:

$$\begin{aligned} L(x_{1:n}; \alpha) &= \prod_{k=1}^{n} \alpha e^{-\alpha x_x} = \alpha^n e^{-\alpha\sum_{k=1}^{n} x_k} \\ &\Rightarrow \\ \tilde{L} &= n\log(\alpha) - \alpha\sum_{k=1}^{n} x_k \end{aligned}$$

Set the derivative to zero:

$$\frac{d\tilde{L}}{d\alpha} = 0$$

$$0 = \frac{n}{\alpha} - \sum_{k=1}^{n} x_k \Rightarrow \hat{\alpha}_{ML} = \frac{n}{\sum_{k=1}^{n} x_k} = \frac{1}{\bar{x}}$$

(iii) **(The multinomial distribution)** Sample n times, independently, from a distribution on $\{1, 2, \ldots, m\}$, where the probability of getting j on any given sample is p_j. Assume that these probabilities are unknown. Let $X_k = x_k \in \{1, \ldots, m\}$ be the outcome of the kth sample, $k = 1, \ldots, n$. Estimate $p \triangleq (p_1, \ldots, p_m)$.

$$L(x_{1:n}; p) = \prod_{k=1}^{n} p_{x_k} = \prod_{j=1}^{m} p_j^{n_j}$$

where n_j is the number of times that the sample produces j, i.e. $n_j = \#\{k : x_k = j\}$. Hence

$$\tilde{L} = \sum_{j=1}^{m} n_j \log(p_j)$$

To maximize $\tilde{L}$ we need to take into account the fact that $\sum_{j=1}^{m} p_j = 1$. The right way to do this is to use a Lagrange multiplier, λ, and maximize

$$\sum_{j=1}^{m} n_j \log(p_j) + \lambda(1 - \sum_{j=1}^{m} p_j)$$

instead of just $\tilde{L}$. Taking partial derivatives, and setting them equal to zero, we get $\frac{n_j}{p_j} - \lambda = 0$ for all $j = 1, \ldots, m$. Hence $p_j = \frac{n_j}{\lambda}$, where we need to choose λ to satisfy the constraint, $\sum_{j=1}^{m} p_j = 1$. This works with $\lambda = n$ and we end up with the intuitive estimator $\hat{p}_j = \frac{n_j}{n}$ for each j.

(iv) **(The uniform distribution)** Assume that $X_{1:n} \sim \ iid\ U[a, b]$, for unknown a and b with $a < b$. In this case f is a pdf:

$$f(x; a, b) = \frac{1}{b - a} \mathbb{1}_{x \in [a,b]}$$

and the likelihood is

$$L(x_{1:n}; a, b) = \prod_{k=1}^{n} \frac{1}{b - a} \mathbb{1}_{x_k \in [a,b]} = \begin{cases} \left(\frac{1}{b-a}\right)^n & \text{if } a \leq x_k \leq b \ \forall\ k = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

We want to make a as big as possible and b as small as possible, but without violating $x_k \in [a, b]$ for every k. By inspection

$$\hat{a}_{ML} = \min\{x_k : 1 \leq k \leq n\} \quad \text{and} \quad \hat{b}_{ML} = \max\{x_k : 1 \leq k \leq n\}$$

(v) **(The exponential families–optional)**

Definition. $f_X(x;\theta) = f_X(x;\theta_1,\ldots,\theta_m)$ *is an* exponential family *if it can be written in the form*

$$f_X(x;\theta) = \frac{1}{Z_\theta} h(x) e^{\sum_{i=1}^m \theta_i S_i(x)} \tag{18}$$

Where Z_θ is a normalizing constant, ensuring that $\int f_X = 1$, $h(x)$ is any positive function of x, and $S_i(x)$, $i = 1,\ldots,m$ are (essentially) arbitrary functions, known as "sufficient statistics."

Given $X_{1:n}$ an iid sample from f_X, with values $X_k = x_k$, $k = 1 \ldots m$,

$$L = \prod_{k=1}^n \frac{1}{Z_\theta} h(x) e^{\sum_{i=1}^m \theta_i S_i(x_k)}$$

and

$$\tilde{L} = -n \log(Z_\theta) + \sum_{k=1}^n \log h(x_k) + \sum_{k=1}^n \sum_{i=1}^m \theta_i S_i(x_k)$$

If we take the partial derivative with respect to θ_j and set it equal to zero, then after a lot of algebra (the function Z_θ complicates everything) we arrive at "the likelihood equations"

$$E_\theta[S_i(X)|\theta_1,\ldots,\theta_m] = E_{\hat{F}}[S_i(X)] = \frac{1}{n}\sum_{k=1}^n S_i(x_x) \quad j = 1,\ldots,m$$

Here, E_θ refers to expectation with respect to $f_X(x,\theta)$, whether or not θ is the true sampling distribution, and $\hat{F}$ is the empirical distribution. The likelihood equations make a lot of sense: we choose the unknown parameters to make the expected values of the sufficient statistics equal to the empirical values. The functions $S_i(x)$ are "sufficient" in the sense that knowing their values at each of the samples is enough to determine the maximum-likelihood estimators for all of the unknown parameters.

Most (but not all) of the standard distributions that come up in practice are members of exponential families. There are very general results about estimation, optimality, and consistency that have been proven for the exponential families, and hence inherited by all of the many distributions that are members of exponential families.

Let's go back to the estimation of the mean and variance of a normal distribution, which we worked through earlier. Do normal densities have the right form to be members of exponential families? In other words, can we manipulate $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(x-\mu)^2/2\sigma^2\right)$ into the form specified by (18)?

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x}$$

which fits the bill, with $\theta_1 = \frac{1}{2\sigma^2}$ associated with the statistic $S_1(x) = -x^2$ and $\theta_2 = \frac{\mu}{\sigma^2}$ associated with the statistic $S_2(x) = x$. Accordingly, the likelihood equations are

$$E_{\mu,\sigma^2}[-X^2] = -\frac{1}{n}\sum_{k=1}^n x_k^2 \tag{19}$$

$$E_{\mu,\sigma^2}[X] = \frac{1}{n}\sum_{k=1}^n x_k = \bar{x} \tag{20}$$

But since $E[-X^2] = -(\sigma^2 + \mu^2)$ and $E[X] = \mu$, the likelihood equations can be written as

$$\sigma^2 + \mu^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2$$
$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{x}$$

from which it follows that $\hat{\mu}_{ML} = \bar{x}$ and $\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \bar{x}^2$, in agreement with what we derived earlier.

Remember that not all distributions are in an exponential family. Two notable exceptions are the Cauchy and uniform distributions.

3.5.3 Consistency and KL Divergence

Under mild conditions, maximum-likelihood estimators have a number of desirable properties. Here are three of the most important. (I will use θ_o, generically, to denote the true, and of course unknown, value of the parameter being estimated.)

1. **Consistency.** There is no need to separately prove consistency of each MLE estimator. The method guarantees that $\hat{\theta}_{ML} \to \theta_o$ as $n \to \infty$, e.g. in the MSE sense. Hence both the bias and the variance go to zero.

2. **Asymptotic Normality.** This is quite handy. Whatever the form of the estimator, the shape of its distribution is increasingly normal as the sample size grows, merely because it is an MLE. Formally, $(\hat{\theta}_{ML} - \theta)/\sqrt{\mathbb{V}[\hat{\theta}_{ML}]}$ converges in distribution to $\mathcal{N}(0,1)$ as $n \to \infty$.

3. **Efficiency.** The MLE is asymptotically optimal, meaning roughly that the MSE of the maximum-likelihood estimator goes to zero at least as fast as the MSE of any other consistent estimator.

(The following discussion of the KL divergence, and its role in proving the consistency of the MLE, is optional.)

Let's look more closely at the first property, consistency. The key quantity in studying consistency of maximum likelihood is the *Kullback-Leibler divergence*, which is a general and very powerful measure of the difference between two distributions. Let's look at the case of continuous distributions (definitions for other kinds of distributions are similar):

Definition. *Given two probability density functions, f and g, the* Kullback-Leibler Diveregence*, denoted $D(f||g)$, is defined by*

$$D(f||g) \triangleq \int_{x=-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx$$

The justification for calling it a measure of the difference between f and g is that for any pair of densities: (i) $D(f||g) \geq 0$, and (ii) $D(f||g) = 0$ if and only if f and g are the same density.

The KL divergence comes up in many fields, including probability theory (in the treatment of so-called large deviations), in information theory (as a measure of the number of extra bits you pay when encoding a signal under the assumption that the signal is generated by g when in fact it was generated by f), and in the theory of maximum likelihood, as follows:

Given $X_1, \ldots, X_n \sim \ iid\ f(x;\theta_o)$ (θ_o is the true value of the unknown parameter θ), let $\hat{\theta}_{ML}$ be the maximum-likelihood estimator. In order to emphasize the dependency of the estimator on the sample, and hence it's random character, I will write everyting in terms of $X_{1:n}$ rather than $x_{1:n}$:

$$
\begin{aligned}
\hat{\theta}_{ML} &\triangleq \operatorname*{argmax}_\theta L(X_{1:n};\theta) \\
&= \operatorname*{argmax}_\theta \prod_{k=1}^{n} f(X_k;\theta) \\
&= \operatorname*{argmax}_\theta \sum_{k=1}^{n} \log f(X_k;\theta) \\
&= \operatorname*{argmax}_\theta \frac{1}{n}\sum_{k=1}^{n} \log f(X_k;\theta) \\
&\approx \operatorname*{argmax}_\theta E_{\theta_o}\left[\log f(X;\theta)\right] \quad \text{(law of large numbers, wrt } f(x;\theta_o)\text{)} \\
&= \operatorname*{argmax}_\theta \int_{-\infty}^{\infty} f(x;\theta_o)\log f(x;\theta)dx - \int_{-\infty}^{\infty} f(x;\theta_o)\log f(x;\theta_o)dx \\
&\text{(since subtracting a constant doesn't change "argmax")} \\
&= \operatorname*{argmax}_\theta \int_{-\infty}^{\infty} f(x;\theta_o)\log \frac{f(x;\theta)}{f(x;\theta_o)}dx \quad \text{(know your logarithms!)} \\
&= \operatorname*{argmin}_\theta -\int_{-\infty}^{\infty} f(x;\theta_o)\log \frac{f(x;\theta)}{f(x;\theta_o)}dx \\
&= \operatorname*{argmin}_\theta \int_{-\infty}^{\infty} f(x;\theta_o)\log \frac{f(x;\theta_o)}{f(x;\theta)}dx \\
&= \operatorname*{argmin}_\theta D(f(x;\theta_o)||f(x;\theta)) = \theta_o
\end{aligned}
$$

There is only one approximation, which uses the law of large numbers. So it is reasonable to expect that when n is sufficiently large, $\hat{\theta}_{ML} \approx \theta_o$. In other words, maximizing likelihood is approximately the same thing as minimizing KL divergence between the true density and the estimated density. Given the properties of the KL divergence, the minimizer, $\hat{\theta}_{ML}$, is the value of θ that makes $f(x;\theta) \approx f(x;\theta_o)$, which in general means that θ is close to θ_o.

3.5.4 Failure Modes

Maximum likelihood can fail. Perhaps the most common failure, though no fault of the *principle*, comes from the typical high complexity of likelihood surfaces when $x \in \mathbb{R}^d$, even when d is only a few dimensions, much less a few dozen. There are general optimization methods that can help, but in fact it is typically a game of iterative guessing, as in running a gradient ascent, repeatedly, from many starting positions. In such cases, it is not reasonable to expect to find the *global* maximum of the likelihood.

In addition to computational problems, there are other, more fundamental, problems that can arise:

Non-identifiability. Let's start with *identifiability*:

Definition. *Let $\{P_\theta\}_{\theta\in\Theta}$ be a family of distributions. The parameter θ is said to be **identifiable** if $P_\theta = P_{\theta'} \implies \theta = \theta'$.*

Examples include all of the finite-dimensional maximum-likelihood problems that we worked through earler.

For a more-or-less transparent example of *non-identifiability*, suppose that $X_1, \ldots, X_n \sim \ iid\, \mathcal{N}(\mu_1 + \mu_2, 1)$. No amount of data is going to separate μ_1 from μ_2. To see this, note, for example, that $\mu_1 = \mu_2 = 0$ is indistinguishable from μ_1 =1,000,000 and μ_2 =-1,000,000: If $X \sim \mathcal{N}(\mu_1+\mu_2, 1)$ with $\mu_1 = \mu_2 = 0$ and $Y \sim \mathcal{N}(\mu_1 + \mu_2, 1)$ with μ_1 =1,000,000 and μ_2 =-1,000,000, then X and Y have identical distributions. Of course we may, and often do, care more about the *distribution* of the data than the values of the parameters, per se. In this case, if the mean really is zero, then μ_1 =1,000,000 and μ_2 =-1,000,000 is just as good as $\mu_1 = \mu_2 = 0$.

A close cousin of non-identifiability is near-non-identifiability, in which case the likelihood surface will be very flat in and around its maximum. A gradient-type algorithm for finding the MLE can spend a lot of computing cycles making tiny incremental improvements in the likelihood while still being very far from the maximum.

Miss-specification. More serious and more common is the problem of model miss-specification. Earlier in the course I argued that all models are wrong, and I will stick to that, but the consequences can be, well, inconsequential or catastrophic. What happens when we estimate parameters under an incorrect model? Financial and other types of data, including pixel data, tends to have "heavy tails," meaning that the probabilities of large deviations away from the norm can be much higher than you would expect of, say, a Gaussian distribution. These tail events are what Nassim Taleb, in his book *The Black Swan: The Impact of the Highly Improbable*, calls "black swans." The book is really about model miss-specification. In particular, the Gaussian type models used so often in financial models (e.g. the Black-Scholes model for option pricing) appear to systematically underestimate the probabilities of large deviations: there are "too many" rare events (black swans), at least too many if we are to believe the models.

Let $X_1 \cdots X_n$ i.i.d $\sim f_\mu$, where f_μ is an *unknown density* which is symmetric about the point μ. That is, $f_\mu(\mu + x) = f_\mu(\mu - x)$ for all x.

Given only this information, we might decide to assume that $X_1 \cdots X_n \sim \mathcal{N}(\mu, \sigma^2)$. This will yield the maximum likelihood estimator:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

Unfortunately, this estimator is not always consistent for μ. For example, the Cauchy density is symmetric around μ:

$$f_\mu(x) = \frac{1}{\pi\left(1 + (x-\mu)^2\right)}$$

However, if $X_1 \cdots X_n$ i.i.d $\sim f_\mu$, then $\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k$ will never converge. Indeed, the variance of $\hat{\mu}$ does not even decrease as $n \to \infty$. This remarkable fact is possible because the variance of $\hat{\mu}$ is actually infinite.[9]

Infinities. Sometimes there just isn't *any* maximum-likelihood estimator. No matter how big you make the likelihood, you can always find parameters that will make it even bigger. The circumstances can be exotic (e.g. infinite-dimensional) or mundane (e.g. a mixture of uniform random variables). Check out Assignment 5 for some examples.

[9] All is not lost. Although the Cauchy distribution is not in an exponential family, there are consistent (and even well-behaved) estimators for μ, e.g. $\hat{\mu} = \text{median}(X_1 \cdots X_n)$.