Notes on Probability

Alex Nelson[*]
Email: pqnelson@gmail.com

June 21, 2015

Contents

[*] This is a page from http://pqnelson.github.io
Compiled: June 21, 2015 at 5:10pm (PST)

1 Introduction

1.1. Definition. The set of all possible outcomes of an experiment is called the "**Sample Space**" and denoted Ω. An "**Elementary Event**" is an element of Ω, whereas a "**Event**" is a subset of Ω.

CAUTION: We will abuse notation, and mix up the singleton $\{x\}$ with the element x. So $\{x\}$ is an elementary event, and usually just referred to as x.

1.2. Example. Flipping a coin has two results: heads H, tails T. The sample space is

$$\Omega = \{\, H, T \,\}. \tag{1.1}$$

What's an event? Well, lets consider a few:

1. We flip the coin and get a heads.
2. We get either a heads or a tails.
3. The outcome is both a heads and tails.
4. The outcome is not a heads.

Note that the first and last examples are elementary events, the others are not elementary.

1.3. Remark. This process "flipping a coin", is generalized in mathematics to any experiment with two outcomes: either heads or tails; the baby is either a boy or a girl; the cat is either dead or alive[1]. This experiment is called a "**Bernoulli trial**", and it's the foundation of most (all?) of probability theory.

1.4. Example. Not all sample spaces are finite. For example, consider an experiment describing the decay of an unstable particle. How long does it take? Well, the sample space would be

$$\Omega = \{x \in \mathbb{R} : x \geq 0\}. \tag{1.2}$$

This is quite infinite!

1.5. Definition. We want to think of subsets of the sample space as *events*. The sample space is a "certain event": something's *definitely* going to happen. So now we want to define the "collection of all events (of our sample space)"... but not every subset is an event! So we need some axioms/specifications.

We define a "**σ-Field**" (or *σ-Algebra*) $\mathcal{F}$ to be the set of events of our sample space Ω. But that's not the end of the story: we have a bunch of axioms to consider.

First, it seems sound to suggest for any pair of events A and B (i.e., $A, B \in \mathcal{F}$), we can form new events "A and B" as well as "A or B". These correspond to the operations

$$A \text{ and } B = A \cap B, \quad \text{and} \quad A \text{ or } B = A \cup B.$$

Good, well, so what?

Axiom (Closed under pair-wise "And", "Or"). If $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$.

Under a similar vein of reasoning, if we have an event $A \in \mathcal{F}$, then its complement A^{C} (read "The event that A does not occur") should also be an event: $A^{\mathsf{C}} \in \mathcal{F}$. So, we have

[1] When observed!

Axiom (Closed under complements). If $A\in\mathcal{F}$, then $A^{\complement}\in\mathcal{F}$.

The last axiom is quite simple: nothing is an event. What's "nothing"? The empty set:

Axiom (Nothing is an event). We have $\emptyset\in\mathcal{F}$.

Is this really the last axiom? No, we weren't honest with our first axiom. We have something *more*: we could have an infinite number of "and" (but not an infinite number of "or").

Axiom. If $A_i\in\mathcal{F}$, then $\displaystyle\bigcup_{i=1}^{\infty}A_i\in\mathcal{F}$

When is this useful? Suppose we want to flip a coin, and keep flipping until we get a heads. What's the sample space look like? Well, it'd be

$$\Omega=\{\,H,TH,TTH,TTTH,\dots\}.\tag{1.3}$$

The event that we flip the coin an even number of times is

$$E=\{\,TH,TTTH,TTTTTH,\dots\}.\tag{1.4}$$

Unless we have this last axiom, we couldn't construct it!

1.6. Example. The smallest σ-algebra associated to any sample space Ω is

$$\mathcal{F}=\{\emptyset,\Omega\}.\tag{1.5}$$

It "obviously" satisfies the four axioms.

1.7. Example. The next smallest algebra associated to Ω is, if A is any subset of Ω, then

$$\mathcal{F}=\{\,\emptyset,A,A^{\complement},\Omega\,\}.\tag{1.6}$$

Although a little trickier to show, it also satisfies the axioms.

1.8. Example. When Ω is finite[2], its powerset (the set of all subsets) of Ω is a σ-algebra. This is the most common σ-algebra used when working with finite sample spaces.

2 Probability

2.1. There are two interpretations to probability. Avoiding philosophical arguments, we suggest probability is a function $\Pr(-)$ that assigns to each event in our sample space $X\in\Omega$ a number $\Pr(X)$ such that a bunch of axioms hold.

2.1.1. Remark. The "Objectivist" interpretation suggests

$$\Pr(X)=\frac{\text{number of trials where }X\text{ is the value}}{\text{total number of trials}}=\frac{N(X)}{N(\Omega)}\tag{2.1}$$

whereas Bayesian probability theorists suggest probability is really "belief" that X is the outcome of a trial. In either event, we can deduce a number of axioms from Equation (2.1).

[2] For *infinite* sample spaces, things get tricky because we're really going to do "integration" on our set. For the real numbers, for example, its powerset includes the natural numbers... but an integral over the natural numbers embedded in the reals is zero! We get strange results like that: where things should have some probability, they instead have none.

2.2. Probability. First we should note we cannot have more outcomes taking value X than there are outcomes:

$$N(X) \leq N(\Omega) \tag{2.2}$$

Consequently, when we divide through both sides we get

$$\frac{N(X)}{N(\Omega)} \leq 1. \tag{2.3}$$

Similarly cannot have a "negative number" of events occur, so

$$0 \leq N(X) \tag{2.4}$$

for any X. Thus we see

$$0 \leq \Pr(X) \leq 1 \quad \text{for any } X \in \Omega. \tag{2.5}$$

This is one axiom.

Axiom. We have $\Pr(X) \in [0,1]$, i.e., $0 \leq \Pr(X) \leq 1$ for any $X \in \mathcal{F}$.

Observe, we can deduce from this another specification. Namely,

$$\Pr(\Omega) = \frac{N(\Omega)}{N(\Omega)} = 1. \tag{2.6}$$

Similar reasoning suggests

$$\Pr(\emptyset) = \frac{0}{N(\Omega)} = 0. \tag{2.7}$$

These form another axiom.

Axiom (Certainty Something Happens, Nothing Happens). We have $\Pr(\Omega) = 1$ and $\Pr(\emptyset) = 0$.

When we consider events $X_1, \dots, X_n \in \mathcal{F}$ which are disjoint (so $X_i \cap X_j = \emptyset$ for $i \neq j$), what happens to the probability? We have

$$\Pr\left(\bigcup_i X_i\right) = \sum_i \Pr(X_i).$$

Does this make sense? The intuition should be "The probability that one of the X_i's occur is the sum of the probability of each event" which makes sense if they're disjoint events (there's "no overlap"). This gives us our last axiom:

Axiom (Disjoint Events). If $X_i \in \mathcal{F}$ is a (possibly infinite) family of disjoint events, then

$$\Pr\left(\bigcup_i X_i\right) = \sum_i \Pr(X_i). \tag{2.8}$$

2.1 Examples

2.3. Coin Tossing. We flip a coin once. The coin may be biased or fair. We take $\Omega = \{H, T\}$ and $\mathcal{F} = \{\emptyset, H, T, \Omega\}$. A possible probability measure[3]

$$\Pr\colon \mathcal{F} \to [0,1]$$

given by

$$\Pr(H) = p, \quad \Pr(T) = 1 - p \tag{2.9}$$

where $0 \leq p \leq 1$, and the "obvious values" $\Pr(\emptyset) = 0$, $\Pr(\Omega) = 1$. When $p = 1/2$, then we call the coin "fair" or "unbiased."

[3] There are many different acceptable probability measures, if we are being honest. But this is the measure the reader probably has in mind.

2.4. Dice. A six-sided die is thrown once. We have the possible outcomes be $\Omega = \{1,2,3,4,5,6\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ (where $\mathcal{P}(X)$ is the power set of X). The probability measure Pr is given by

$$\Pr(A) = \sum_{i\in A} p_i \quad \text{for any } A \subseteq \Omega \tag{2.10}$$

where $p_1, \ldots, p_6$ are specified numbers in the unit interval $[0,1]$ whose sum is 1. Note that p_k is the probability we roll a k. The die is "fair" if

$$\Pr(A) = \frac{|A|}{6} \quad \text{for any } A \subseteq \Omega \tag{2.11}$$

where $|A|$ is the Cardinality of A.

2.5. Definition. A "**Probability Space**" consists of a sample space Ω equipped with its σ-algebra $\mathcal{F}$ and probability measure $\Pr\colon \mathcal{F} \to [0,1]$.

We will often simply state "Given a probability space $(\Omega, \mathcal{F}, \Pr)$, ..." with the understanding what each component means.

Note that a probability space represents one experiment. It's the outcomes (i.e., the sample space) equipped with the events (i.e., the σ-algebra $\mathcal{F}$) and a description of outcomes (i.e., probability measure).

2.6. Lemma. *Given a probability space $(\Omega, \mathcal{F}, \Pr)$, and $A, B \in \mathcal{F}$, then the following hold:*

1. *For any $A \in \mathcal{F}$, we have $\Pr(A^{\complement}) = 1 - \Pr(A)$.*
2. *If $A \subseteq B$, then $\Pr(B) = \Pr(A) + \Pr(B \setminus A) \geq \Pr(A)$.*
3. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Proof. (1) We see that $A^{\complement} \cup A = \Omega$ and $A^{\complement} \cap A = \emptyset$. So these are disjoint events, and by our axioms we have $\Pr(A \cup A^{\complement}) = \Pr(A) + \Pr(A^{\complement}) = 1$. Thus $\Pr(A^{\complement}) = 1 - \Pr(A)$.

(2) We see that $A \cap (B \setminus A) = \emptyset$, which implies

$$\Pr\big(A \cup (B \setminus A)\big) = \Pr(A) + \Pr(B \setminus A).$$

But $A \cup (B \setminus A) = B$, which implies the result.

(3) We see $A \cup B = A \cup (B \setminus A)$. Then

$$\Pr(A \cup B) = \Pr(A) + \Pr(B \setminus A)$$

since the right hand side is disjoint. We then note that

$$B \setminus A = B \setminus (A \cap B)$$

which allows us to write

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B \setminus A)\\ &= \Pr(A) + \Pr\big(B \setminus (A \cap B)\big)\end{aligned}$$

and using result (2) we have

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \tag{2.12}$$

as desired. □

2.7. Lemma. *For $A_1, \dots, A_n$ events (not necessarily disjoint), we have*

$$\begin{aligned}\Pr\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_i \Pr(A_i) - \sum_{i<j} \Pr(A_i \cap A_j) \\ &\quad + \sum_{i<j<k} \Pr(A_i \cap A_j \cap A_k) + \dots \\ &\quad + (-1)^{n+1} \Pr(A_1 \cap \dots \cap A_n).\end{aligned} \tag{2.13}$$

Note this is a more general result than lemma 2.6's. We prove it by induction.

Proof. **Base Case ($n = 2$):** this is precisely lemma 2.6's result.

Inductive Hypothesis: assume this works for arbitrary n.

Inductive Case: When we have $A_1, \dots, A_n, A_{n+1}$, we have

$$\Pr\left(\bigcup_{i=1}^{n+1} A_i\right) = \Pr\left(\bigcup_{i=1}^{n} A_i \cup A_{n+1}\right) \tag{2.14}$$

Let $B = \bigcup_{i=1}^{n} A_i$, then we rewrite this equation as

$$\Pr\left(\bigcup_{i=1}^{n+1} A_i\right) = \Pr\left(B \cup A_{n+1}\right) \tag{2.15}$$

which is *precisely* the base case! □

2.8. Proposition. Let A_k be a sequence of increasing events, i.e.,

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots \tag{2.16}$$

Let

$$A = \bigcup_{k=1}^{\infty} A_k = \lim_{k\to\infty} A_k, \tag{2.17}$$

then

$$\Pr(A) = \lim_{k\to\infty} \Pr(A_k). \tag{2.18}$$

Similarly, if B_j is a decreasing sequence of events, so $B_1 \supseteq B_2 \supseteq B_3 \supseteq \dots$, then

$$B = \bigcap_{j=1}^{\infty} B_j = \lim_{j\to\infty} B_j \tag{2.19}$$

satisfies

$$\Pr(B) = \lim_{j\to\infty} \Pr(B_j). \tag{2.20}$$

Proof. For the first statement, it's easy to see

$$\bigcup_{k=1}^{N} A_k \subseteq A_N \tag{2.21}$$

We can take the limit as $N \to \infty$ on both sides to get the desired relation. We also see that

$$A_1 \cup \bigcup_{k=2}^{N} (A_k \setminus A_{k-1}) = \bigcup_{k=1}^{N} A_k. \tag{2.22}$$

Thus we find

$$\Pr(A) = \Pr(A_1) + \lim_{N\to\infty}\sum_{k=2}^{N}\Pr(A_k) - \Pr(A_{k-1}) = \lim_{N\to\infty}\Pr(A_N). \tag{2.23}$$

Similar reasoning holds for the second statement. □

2.2 Conditional Probability

2.9. Suppose we have two events A, B. What's the probability, if B occurs, that A will occur? These sort of conditional statements we're interested in, usually in scientific fields. What would the probability look like? Lets denote $\Pr(A|B)$ be the probability of A given B. Then the probability of A and B happening would be

$$\Pr(A|B)\Pr(B) = \Pr(A\cap B). \tag{2.24}$$

If we accept this, then

$$\Pr(A|B) = \frac{N(A\cap B)}{N(B)} = \frac{\Pr(A\cap B)}{\Pr(B)} \tag{2.25}$$

where we implicitly divide both the numerator and denominator by $N(\Omega)$ to get the fraction of probabilities.

2.10. Definition. If $\Pr(B) > 0$, then the "**Conditional Probability**" that A occurs given B definitely occurs is defined as

$$\Pr(A|B) = \frac{\Pr(A\cap B)}{\Pr(B)}. \tag{2.26}$$

2.11. Example (Children). Suppose a couple has two children. The sample space is

$$\Omega = \{\, BB, BG, GB, GG \,\} \tag{2.27}$$

where the each element of the sample space indicates what the children are (so BG indicates the first is a boy, while the second is a girl). What's the probability, given one child is a boy, that both children are boys?

Solution: Well, we see that the event one is a boy X is really

$$X = \{\, BG, GB, BB \,\} \tag{2.28}$$

We suppose for simplicity that the probability of each outcome in the sample space is equal, so

$$\Pr(GG) = \Pr(GB) = \Pr(BG) = \Pr(BB) = 1/4. \tag{2.29}$$

Thus we see

$$\begin{aligned}\Pr(BB|X) &= \frac{\Pr(BB\cap X)}{\Pr(X)}\\ &= \frac{\Pr(BB)}{\Pr(X)}\\ &= \frac{1/4}{3/4} = \frac{1}{3}.\end{aligned} \tag{2.30}$$

Note this is contrary to popular intuition, which would vaguely suggest the solution is $1/4$.

2.12. Lemma. *Let A and B be events, $0 < \Pr(B) < 1$. Then*

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|B^{\mathsf{C}})\Pr(B^{\mathsf{C}}). \tag{2.31}$$

The proof is direct.

Proof. We substitute the definition of conditional probability

$$\begin{aligned}
\Pr(A) &= \Pr(A|B)\Pr(B) + \Pr(A|B^{\mathsf{C}})\Pr(B^{\mathsf{C}}) && \text{(2.32a)}\\
&= \frac{\Pr(A\cap B)}{\Pr(B)}\Pr(B) + \frac{\Pr(A\cap B^{\mathsf{C}})}{\Pr(B^{\mathsf{C}})}\Pr(B^{\mathsf{C}}) && \text{(2.32b)}\\
&= \Pr(A\cap B) + \Pr(A\cap B^{\mathsf{C}}) && \text{(2.32c)}
\end{aligned}$$

But look, the events $A\cap B$ and $A\cap B^{\mathsf{C}}$ are disjoint. So we have

$$\Pr(A\cap B) + \Pr(A\cap B^{\mathsf{C}}) = \Pr\big((A\cap B)\cup(A\cap B^{\mathsf{C}})\big). \tag{2.32d}$$

Look, this is quite simply

$$\Pr\big((A\cap B)\cup(A\cap B^{\mathsf{C}})\big) = \Pr(A) \tag{2.32e}$$

precisely as desired. □

2.13. Lemma. *Let B_i be a family of disjoint events such that*

$$\bigcup_i B_i = \Omega. \tag{2.33}$$

Then

$$\Pr(A) = \sum_i \Pr(A|B_i)\Pr(B_i). \tag{2.34}$$

The proof is similar to the previous lemma.

Proof. We begin by using the definition of conditional probability

$$\begin{aligned}
\sum_i \Pr(A|B_i)\Pr(B_i) &= \sum_i \frac{\Pr(A\cap B_i)}{\Pr(B_i)}\Pr(B_i) && \text{(2.35a)}\\
&= \sum_i \Pr(A\cap B_i). && \text{(2.35b)}
\end{aligned}$$

But look, the events $A\cap B_i$ are disjoint since B_i is a family of *disjoint* events. So we can write this as

$$\sum_i \Pr(A\cap B_i) = \Pr\left(\bigcup_i A\cap B_i\right) \tag{2.35c}$$

Using set theoretic properties of set union and intersection, we can write this as

$$\Pr\left(\bigcup_i A\cap B_i\right) = \Pr\left(A\cap\bigcup_i B_i\right) \tag{2.35d}$$

By hypothesis, the union of all B_i is the sample space, so we have

$$\Pr\left(A\cap\bigcup_i B_i\right) = \Pr(A\cap\Omega) = \Pr(A) \tag{2.35e}$$

precisely as desired. □

2.14. Remark on Examples: Inheritance from Past. In probability, a lot of examples involve drawing items (e.g., colored balls, slips of paper, etc) from urns. This is because the Bernoulli family, who pioneered most of probability theory in 18th century France, were using then-contemporary situations. You would cast ballots in an urn, etc. Even today in France, the phrase "going to vote" is *aller aux urnes.*

2.15. Example (Balls from Urn). Given two urns, each containing colored balls. Urn I contains two white and three blue balls, urn II contains three white and four blue balls. A ball is drawn randomly from urn I and put into urn II, then a balled is picked at random from urn II and examined. What is the probability the examined ball is blue?

Solution: Really, this is two events going on. Event 1 is transferring a ball from urn I to urn II, and event 2 is the color of the ball drawn from urn II.

So, let B be the event that a white ball is transferred from urn I to urn II. Then B^{C} is the event it's a blue ball transferred. The event A is the examined ball is blue. So

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|B^{\mathsf{C}})\Pr(B^{\mathsf{C}}) \tag{2.36}$$

But look, we can start calculating some stuff out:

$$\Pr(B) = \frac{2}{5}, \quad \text{and} \quad \Pr(B^{\mathsf{C}}) = \frac{3}{5}. \tag{2.37}$$

The conditional probabilities are easier to compute now:

$$\Pr(A|B) = \frac{4b}{4w+4b} = \frac{1}{2} \tag{2.38}$$

where w stands for "white ball", b for "blue ball", and

$$\Pr(A|B^{\mathsf{C}}) = \frac{5b}{3w+5b} = \frac{5}{8}. \tag{2.39}$$

Thus we see

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|B^{\mathsf{C}})\Pr(B^{\mathsf{C}}) \tag{2.40a}$$

$$= \frac{1}{2}\cdot\frac{2}{5} + \frac{5}{8}\cdot\frac{3}{5} = \frac{23}{40}. \tag{2.40b}$$

2.16. Example (Elbonian Widgets). In the tiny country of Elbonia[4], there are two widget factories. But 20% of the widgets produced by factory I are defective, whereas 5% from factory II are defective. Factory I produces twice as many widgets as factory II. What's the probability a given Elbonian widget is satisfactory?

Solution: Let A be the event the widget is satisfactory, and B the event it's from factory I. We see

$$\Pr(B) = \frac{2}{3} \tag{2.41}$$

and so

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|B^{\mathsf{C}})\Pr(B^{\mathsf{C}}). \tag{2.42}$$

The conditional probabilities are given, or at least easily deduced

$$\Pr(A|B) = 1 - \frac{1}{5} \tag{2.43a}$$

$$\Pr(A|B^{\mathsf{C}}) = 1 - \frac{1}{20}. \tag{2.43b}$$

[4] A fictional country from the comic strip "Dilbert."

We see then

$$\begin{aligned}\Pr(A) &= \frac{4}{5}\cdot\frac{2}{3} + \frac{19}{20}\cdot\frac{1}{3}\\ &= \frac{51}{60}.\end{aligned} \tag{2.44}$$

This concludes our example (roughly 5 out of 6 Elbonian widgets are satisfactory).

2.3 Independent Events

2.17. Suppose we have two events X and Y. If Y does not depend on X, we expect

$$\Pr(Y|X) = \Pr(Y). \tag{2.45}$$

Lets try to consider a slightly more general situation. If we multiply both sides of Eq (2.45) by $\Pr(X)$, we get

$$\begin{aligned}\Pr(Y|X)\Pr(X) &= \Pr(X\cap Y)\\ &= \Pr(Y)\Pr(X)\end{aligned} \tag{2.46}$$

which gives us the desired condition for "independence."

2.18. Definition. Let X, Y be events. We call them "**Independent Events**" if and only if

$$\Pr(X\cap Y) = \Pr(X)\Pr(Y). \tag{2.47}$$

More generally, for a family of events X_i, they are independent iff

$$\Pr\left(\bigcap_j X_j\right) = \prod_j \Pr(X_j) \tag{2.48}$$

Why is this a good definition? If X or Y has probability zero, we avoid the risk of dividing by zero. This could not have been avoided using Eq (2.45). But notice our condition for independence implies Eq (2.45)!

Caution: Do not make the rookie mistake thinking, for a family of events X_j, independence holds iff for each $i\neq j$ we have $\Pr(X_i\cap X_j) = \Pr(X_i)\Pr(X_j)$. This is *pairwise independence*, and not necessarily the same as implying the family consists of independent events.

2.19. Example (Pairwise Independence Problems). Suppose we have

$$\Omega = \{abc, acb, cab, cba, cab, bca, bac, aaa, bbb, ccc\} \tag{2.49}$$

and they are all equal probable outcomes. Let A_k be the event the kth letter is a.

We claim $\{A_1, A_2, A_3\}$ is a family of pairwise independent events. Observe each A_k has three events. For example, $A_1 = \{abc, acb, aaa\}$. Then we see

$$\Pr(A_i\cap A_j) = \Pr(aaa) = \frac{1}{9} \tag{2.50}$$

and

$$\Pr(A_i)\Pr(A_j) = \frac{1}{3}\cdot\frac{1}{3} = \frac{1}{9}. \tag{2.51}$$

Thus we see

$$\Pr(A_i\cap A_j) = \Pr(A_i)\Pr(A_j) \tag{2.52}$$

for $i \neq j$. This is the definition of pair-wise independent.

However, observe

$$\Pr(A_1 \cap A_2 \cap A_3) = \Pr(aaa) = \frac{1}{9} \tag{2.53}$$

whereas

$$\Pr(A_1)\Pr(A_2)\Pr(A_3) = \frac{1}{27}. \tag{2.54}$$

So we have

$$\Pr(A_1 \cap A_2 \cap A_3) \neq \Pr(A_1)\Pr(A_2)\Pr(A_3). \tag{2.55}$$

That is to say, our family of pairwise-independent events is not a family of independent events!

2.4 Product Space

2.20. Proposition. Let $\mathcal{F}$, $\mathcal{G}$ be σ-algebras over Ω. Then $\mathcal{F} \cap \mathcal{G}$ is also a σ-algebra over Ω.

More generally, if $\mathcal{F}_i$ is a family of σ-algebras over Ω, then

$$\bigcap_i \mathcal{F}_i = \widetilde{\mathcal{F}} \tag{2.56}$$

is a σ-algebra over Ω.

2.21. Problem: We claim that, if $\mathcal{F}$ and $\mathcal{G}$ are σ-algebras over Ω, then $\mathcal{F} \cup \mathcal{G}$ is *not* a σ-algebra over Ω. No!

But we can *uniquely extend* $\mathcal{F} \cup \mathcal{G}$ to a "smallest" σ-algebra containing both $\mathcal{F}$ and $\mathcal{G}$ as subalgebras. What to do? We simply consider the collection

$$\{\mathcal{H}_i : \mathcal{F} \subseteq \mathcal{H}_i, \quad \text{and} \quad \mathcal{G} \subseteq \mathcal{H}_i\} \tag{2.57}$$

then we construct

$$\bigcap_i \mathcal{H}_i = \mathcal{H}. \tag{2.58}$$

This is the smallest such σ-algebra containing both $\mathcal{F}$ and $\mathcal{G}$.

2.22. Recall we describe an experiment using a probability space. But what if we want to have a "composite" experiment? Say, flip a coin *and* draw a card from a deck. How can we describe this experiment? Let us try to consider it!

We want to combine $(\Omega_1, \mathcal{F}_1, \Pr_1)$ and $(\Omega_2, \mathcal{F}_2, \Pr_2)$. What to do?

First lets construct the sample space. We expect, correctly, that

$$\Omega = \Omega_1 \times \Omega_2 \tag{2.59}$$

is our sample space.

Next the σ-algebra. This is more subtle, and requires some justification (given in our discussion of "completeness"). The set $\mathcal{F}_1 \times \mathcal{F}_2$ *is not* a σ-algebra. But we can construct the smallest σ-algebra containing it! We use this "smallest" σ-algebra. We use the same process outlined in §2.21.

The probability measure is simply $\Pr_{12}(A_1 \times A_2) = \Pr_1(A_1)\Pr_2(A_2)$, where $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$.

2.4.1 Completeness

Caution: For a "first read", this discussion of "completeness" may be skipped. It's only useful when dealing with Ω that's infinitely large (something uncountably infinite like $[0,1]\subseteq\mathbb{R}$).

2.23. Probability is a special form of "measure theory", i.e., probability assigns some "volume" to an event. The "volume" is just the likelihood the event will occur, i.e., the "volume" is the event's probability. There are some subtle measure theoretic topics that needs to be discussed. Consequently, this bit on "completeness" can be skipped on the first read. But the motivation will be given, and should be read.

2.24. Motivation: Suppose we have $\mathbb{R}$ with a Lebesgue measure μ. So $\mu(x)=0$ for any $x\in\mathbb{R}$, the length of a point is zero. Then we can try to naively construct the product space $\mathbb{R}^2$ with the measure $\mu^2(A\times B)=\mu(A)\mu(B)$. This has the merit that for any measurable $A\subseteq\mathbb{R}$ we have

$$\mu^2(\{0\}\times A)=\mu(0)\mu(A)=0 \tag{2.60}$$

but only if A is measurable. What if A is not measurable? Well, we *expect* the result to be the same: zero. But instead we get "This is an undefined question!"

2.25. For probability, this is saying if we have a product probability space $\Omega_1\times\Omega_2$ and some event $A\times B$, when A is a subset of a null event we expect $\Pr(A\times B)=0$. This is a technical condition that appears unclear,

2.5 Worked Examples

2.26. Overview. Tools. We will consider some examples, to solidify understanding of the concepts of probability.

Note that one can get a good handle on approximating probability by taking Ω and considering its elements as equi-probable, i.e., $\Pr(x)=1/N$ where $x\in\Omega$ and $N=|\Omega|$. Then any $A\subseteq\Omega$ has probability $\Pr(A)=|A|/N$. This relieves the reader from determining the σ-algebra and the probability measure on it.

There are three basic techniques we will use:

1. Combinatorics: there are $n!$ ways to permute n objects, and $\binom{n}{r}$ different ways to choose r objects (without replacement) from n possible.

2. Set theory: recall we can partition the sample space, and work with the partitions for probability.

3. Independence: don't forget independent events!

2.27. Example (Poker). Every player is dealt 5 cards. For simplicity, suppose each player receives their 5 cards all at once. What is the probability that you, the first to receive cards, will get four-of-a-kind?

Solution: There are 52 cards in a deck, we partition it into the 13 sets of "kinds" (that is, by rank). One approximate solution would suggest that the probability is

$$\Pr(K)\approx\frac{1}{\binom{13}{1}}=\frac{1}{13}\approx 0.076923 \tag{2.61}$$

We can refine this slightly by considering the events.

Let X be the first card drawn. We will consider strings of the form $XYNYN$ where Y indicates the card is the same rank as X, N indicates a different rank. The desired events are

$$A = \{XNYYY, XYNYY, XYYNY, XYYYN\} \tag{2.62}$$

These are, needless to say, independent events. The probabilities are

$$\Pr(XNYYY) = \frac{48 \cdot 3 \cdot 2 \cdot 1}{51 \cdot 50 \cdot 49 \cdot 48} = \frac{1}{20825}, \tag{2.63}$$

and since all five cards are dealt at once, the other probabilities are the same. So we have

$$\Pr(A) = \frac{1}{4165} \approx 0.00024 \tag{2.64}$$

which is *considerably worse* than our first approximation!

Variations: What if we have n players, and each player is dealt only one card at a time until each player has 5 cards? What's the probability of obtaining a four-of-a-kind?

What if we have k decks? What's the probability getting four-of-a-kind?

2.28. Example. There are three cities: A, B, C. Two roads connect A and B, and two roads connect B and C. In Winter, each road has probability p being closed due to snow. This probability is independent of other roads being closed. What is the probability there is an open road connecting A to C?

Solution: Let X be the event there is an open road connecting A and B, Y be the event there is an open road connecting B and C. Then we want to find $\Pr(X \cap Y)$. Luckily, these events are independent, so we have

$$\Pr(X \cap Y) = \Pr(X)\Pr(Y). \tag{2.65}$$

The probability that there is an open road connecting A and B is simply

$$\Pr(X) = 1 - \Pr(X^{\mathsf{C}}) \tag{2.66a}$$

where X^{C} is the event both roads connecting A and B are closed. Since both roads being closed are independent of each other, we find

$$\Pr(X^{\mathsf{C}}) = p^2 \tag{2.66b}$$

and thus

$$\Pr(X) = 1 - p^2. \tag{2.66c}$$

The probability for Y is the same

$$\Pr(Y) = 1 - p^2. \tag{2.67}$$

Thus the probability that some road is open is

$$\Pr(X \cap Y) = (1 - p^2)^2 = 1 - 2p^2 + p^4. \tag{2.68}$$

For example, if $p = 0.5$, then $\Pr(X \cap Y) = 0.5625$.

Variations: What if the probability that both roads (connecting two given cities) is closed becomes dependent on each other?

What if we let there be n cities: $A_1, \ldots, A_n$, and "neighboring" cities A_{j-1}, A_j are connected by precisely 2 roads. What is the probability there exists an open route (sequence of roads) connecting A_1 to A_n?

What if we let each neighboring cities be connected by $k \in \mathbb{N}$ roads? Or have $f(A_i, A_j) \in \mathbb{N}_0$ roads connecting A_i and A_j? That is, we have a completely general undirected graph, what's the probability that any two given cities are connected?

2.29. Example. A man is saving up to a buy a car. His banker advises the man to take up gambling. The man has k units of money, but he needs N. So he goes to a casino, and plays a game with the following rules: the man flips a coin, if it's head the man gets 1 unit of money, and if it's tails the man pays 1 unit of money. What's the probability the man gets N units of money?

Assume it's a fair coin.

Solution: Let B be the event the man goes bankrupt. Let H be the event the first flip is a heads. We will write $\Pr_k$ to indicate that we are working with the condition the amount of money the man has is k units. We use lemma 2.12 to write

$$\Pr_k(B) = \Pr_k(B|H)\Pr_k(H) + \Pr_k(B|H^{\mathsf{C}})\Pr_k(H^{\mathsf{C}}) \tag{2.69a}$$

but we can note $\Pr_*(H) = \Pr_*(H^{\mathsf{C}}) = 1/2$, thus

$$\Pr_k(B) = \frac{\Pr_k(B|H) + \Pr_k(B|H^{\mathsf{C}})}{2}. \tag{2.69b}$$

Note that $\Pr_k(B|H) = \Pr_{k+1}(B)$ and $\Pr_k(B|H^{\mathsf{C}}) = \Pr_{k-1}(B)$ allows us to write this as

$$\Pr_k(B) = \frac{\Pr_{k+1}(B) + \Pr_{k-1}(B)}{2}. \tag{2.69c}$$

Let $p_k = \Pr_k(B)$, then we have a recurrence relationship

$$p_k = \frac{p_{k+1} + p_{k-1}}{2} \tag{2.69d}$$

where $0 < k < N$. Note that $p_0 = 1$ (if the man has no money, he's definitely bankrupt), and $p_N = 0$ (the man stops when he has N units, and hence cannot go bankrupt).

Let $q_k = p_k - p_{k-1}$. We claim

$$q_k = q_{k-1}. \tag{2.70}$$

Subtract $(p_k + p_{k-1})/2$ from both sides of Eq (2.69d) gives us

$$p_k - \frac{1}{2}(p_k + p_{k-1}) = \frac{(p_{k+1} + p_k) - p_k - p_{k-1}}{2} \tag{2.71a}$$

which reduces to

$$\frac{p_k - p_{k-1}}{2} = \frac{p_{k+1} - p_k}{2} \tag{2.71b}$$

or equivalently

$$\frac{q_k}{2} = \frac{q_{k+1}}{2} \tag{2.71c}$$

which proves the claim.

Thus $q_k = q_1$ for any k. But more importantly

$$p_k = q_k + q_{k-1} + \cdots + q_1 + p_0 = kq_1 + p_0. \tag{2.72}$$

Thus we have, using our conditions $p_0 = 1$ and $p_N = 0$

$$q = \frac{-1}{N}, \quad \text{and} \quad p_k = 1 - \frac{k}{N}. \tag{2.73}$$

What does that mean? As the price increases (i.e., as $N \to \infty$), the probability of bankruptcy $p_k \to 1$.

The moral of the story is, of course, avoid bankers.

Variations: What's the probability of bankruptcy if the man uses a biased coin?

2.30. Example (Court). The court investigates whether some event X has happened. There are two witnesses, Tisias and Corax. Given some event, Tisias describes it reliably with probability τ. Likewise, Corax is reliable with probability γ. Let T be the event Tisias asserts X happened, C be the event Corax asserts X happened. Assuming the events C and T are independent (i.e., no collusion between Tisias and Corax), and given both testify X occurred, what is the probability that X really did occur?

Solution: Let $\Pr(X) = x$. Then we note the reliability condition for Corax implies

$$\Pr(C|X) = \gamma, \quad \text{and} \quad \Pr(C|X^{\mathsf{C}}) = 1 - \gamma. \tag{2.74}$$

Then

$$\begin{aligned}\Pr(C) &= \Pr(C|X)\Pr(X) + \Pr(C|X^{\mathsf{C}})\Pr(X^{\mathsf{C}}) &\text{(2.75a)}\\ &= \gamma\cdot x + (1-\gamma)(1-x) &\text{(2.75b)}\\ &= 1 - \gamma - x + 2\gamma x. &\text{(2.75c)}\end{aligned}$$

A similar expression holds for $\Pr(T)$.

Observe, we want to find

$$P(X|T\cap C) = ??? \tag{2.76}$$

We need to find $\Pr(X\cap T\cap C)$ and $\Pr(T\cap C)$. Since T and C are independent ("no collusion"), we have

$$\Pr(T\cap C) = \Pr(T)\Pr(C) \tag{2.77}$$

and similarly

$$\begin{aligned}\Pr(X\cap T\cap C) &= \Pr\big((X\cap T)\cap(X\cap C)\big)\\ &= \Pr(X\cap T)\Pr(X\cap C).\end{aligned} \tag{2.78}$$

which implies

$$\Pr(X|T\cap C) = \frac{\Pr(X\cap T)\Pr(X\cap C)}{\Pr(T)\Pr(C)} = \Pr(X|T)\Pr(X|C). \tag{2.79}$$

Wonderful.

We have to find $\Pr(X|C)$. We see

$$\Pr(X|C) = \frac{\Pr(X\cap C)}{\Pr(C)} \tag{2.80a}$$

by definition, and

$$\Pr(X\cap C) = \Pr(C|X)\Pr(X) = \gamma x. \tag{2.80b}$$

Thus we have

$$\Pr(X|C) = \frac{\gamma x}{1-\gamma-x+2\gamma x} \tag{2.80c}$$

which is one part of the puzzle. (A similar expression holds for Tisias.)

We conclude by writing the entire probability out

$$\begin{aligned}\Pr(X|C\cap T) &= \Pr(X|C)\Pr(X|T)\\ &= \frac{\gamma x}{1-\gamma-x+2\gamma x}\cdot\frac{\tau x}{1-\tau-x+2\tau x}\end{aligned} \tag{2.81}$$

which is completely different than the expected solution!

So, if $\gamma = \tau = 9/10$ and $x = 1/1000$, then

$$\Pr(X|C\cap T) = \left(\frac{9}{1017}\right)^2 \approx 0.00007831466. \tag{2.82}$$

But, on the other hand, the probability that Corax would say X has happened would be

$$\Pr(C) = \frac{1017}{10^4} = 0.1017, \tag{2.83}$$

which is astoundingly high. The moral of *this* story is: don't go to court.

Alternate Solution: It would actually be easier to write up the σ-algebra. We will write the elements of the sample space as triples (a, b, c) where a is 1 if X occurs, 0 otherwise; b is 1 if Tisias asserts X occurred, 0 otherwise; c is 1 if Corax asserts X has occurred, 0 otherwise. So $\Omega = {\mathbb{Z}_2}^3$. Then the relevant events are

$$X = \{(1,0,0), (1,0,1), (1,1,0), (1,1,1)\} \tag{2.84a}$$
$$C = \{(0,0,1), (0,1,1), (1,0,1), (1,1,1)\} \tag{2.84b}$$
$$T = \{(0,1,0), (0,1,1), (1,1,0), (1,1,1)\}. \tag{2.84c}$$

We find, moreover, that

$$\begin{aligned}\Pr(C) &= (1-x)\tau(1-\gamma) + (1-x)(1-\tau)(1-\gamma) + x(1-\tau)\gamma + x\tau\gamma\\ &= \gamma\end{aligned} \tag{2.85}$$

which we expect. We also find

$$\begin{aligned}\Pr(C\cap T) &= (1-x)(1-\gamma)(1-\tau) + x\gamma\tau\\ &= 1 - x - \gamma - \tau + x\gamma + x\tau + \gamma\tau.\end{aligned} \tag{2.86}$$

We then find

$$\Pr(X\cap T\cap C) = x\gamma\tau \tag{2.87}$$

and thus

$$\Pr(X|C\cap T) = \frac{x\gamma\tau}{1 - x - \gamma - \tau + x\gamma + x\tau + \gamma\tau} \tag{2.88}$$

When we set $\gamma = \tau = 9/10$ and $x = 1/1000$ we find $\Pr(X|C\cap T) = 81/1080$. Although slim, it's 75 times greater than if Tisias and Corax said nothing.

2.31. Example. What's the probability that flipping a coin an infinite number of times will result in a heads "sooner or later"?

Solution: We begin by considering the events

$$A_1 = H, \quad A_2 = TH, \quad A_3 = TTH, \tag{2.89}$$

etc. So A_n has $n-1$ tails followed by a heads. Write A for the event that a heads turns up "sooner or later". Then

$$\Pr(A) = \sum_{n=1}^{\infty} \Pr(A_n). \tag{2.90}$$

We see that

$$\Pr(A_1) = 1/2, \tag{2.91a}$$

and

$$\Pr(A_2) = 1/4 \tag{2.91b}$$

or more generally

$$\Pr(A_n) = 2^{-n}. \tag{2.91c}$$

Thus we have

$$\Pr(A) = \sum_{n=1}^{\infty} 2^{-n} = 1. \tag{2.92}$$

So the probability of eventually flipping a heads is 1.

Alternate Solution: Let B_n be the event that we have no heads appear in the first n trials. Then

$$B_1 \supseteq B_2 \supseteq \dots \tag{2.93}$$

So B_1 describes all events starting with 1 tails, B_2 all events starting with 2 tails (which necessarily must include all events starting with 1 tail), and so on.

We use proposition 2.8, let

$$B = \bigcap_{n=1}^{\infty} B_n \tag{2.94}$$

then

$$\Pr(B) = \lim_{n\to\infty} \Pr(B_n). \tag{2.95}$$

But

$$\Pr(B_n) = 2^{-n} \tag{2.96}$$

implies

$$\Pr(B) = \lim_{n\to\infty} 2^{-n} = 0. \tag{2.97}$$

So the probability a head *won't* appear "sooner or later" is 0.

Sample Code. Here's a simple `ANSI C` program which will simply flip a coin and print out the number of tails.

```
#include <time.h>
#include <stdlib.h>
#include <stdio.h>

int main(void)
{
    int n;

    // warm up the random number generator
    srand(time(NULL));
    for(n = 0; n < 100; n++)
        rand();

    // flip the coin
    n = 0;
    while (rand() % 2 == 0)
        printf("Tail number %d\n",++n);

    printf("There were %d tails.\n", n);

    return 0;
}
```

3 Elements of Combinatorial Analysis

3.1. Probability amounts to calculate

$$\Pr(A) = \frac{\begin{pmatrix}\text{number of}\\ \text{elements in } A\end{pmatrix}}{|\Omega|}. \tag{3.1}$$

This is both a blessing and a curse: for we change probability into counting, yet counting can be difficult! We will review elements of combinatorics ("the art of counting").

3.1 Permutations

3.2. Suppose we want to consider how many ways we can order n elements. What to do?

We first say "We have n slots to fill with our n elements." Then we pick some element, and note how many possible slots we may choose. We may pick any one of our n slots. Then pick our second element, and we may place it in any of the remaining $(n-1)$ slots. We continue, and there are

$$n(n-1)(n-2)(\cdots)(1) = n! \tag{3.2}$$

different methods of ordering our n elements. Thus there are $n!$ different permutations.

Convention: We set $0! = 1$.

3.3. Factorial grows quite rapidly. For example

$$3! = 6 \tag{3.3a}$$

$$6! = 720 \tag{3.3b}$$

$$12! = 479\,001\,600 \tag{3.3c}$$

We may approximate it using "**Stirling's Approximation**"

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n. \tag{3.4}$$

We may derive this from the Gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t, \qquad \text{for } \operatorname{Re}(x) > 0. \tag{3.5}$$

We take $\Gamma(n+1)$ and integrate by parts n times, we find

$$\Gamma(n+1) = n! \tag{3.6}$$

for $n \in \mathbb{N}$.

Integrating $\Gamma(x+1)$ by parts merely once will give us the functional equation

$$\Gamma(x+1) = x\Gamma(x). \tag{3.7}$$

Calculus students around the world prove everyday that

$$\Gamma(1/2) = \sqrt{\pi} \tag{3.8}$$

without ever realizing it!

3.2 Sampling

3.4. Given some collection of elements (or a "population"), we would like to pick a finite number of elements (a "sample"). A population with n elements $\{a_1, \ldots, a_n\}$, any *ordered* arrangement of r symbols $a_{j_1}, \ldots, a_{j_r}$ is an "**Ordered Sample of Size** r" drawn from the population. Note when $r = n$, an ordered sample is just a permutation of the population.

Sampling with Replacement: The elements are drawn from the entire population, so we may pick the same element twice (or more).

Sampling Without Replacement: Once we choose an element, we cannot pick it again. We reduce the possible choices, and hence more limited than sampling with replacement.

Choosing r elements with replacement from a population of n can be done in n^r different ways. The proof is obvious: you have r choices, and each choice has n possibilities. So there are

$$\underbrace{n \cdot (\cdots) \cdot n}_{r \text{ times}} = n^r \tag{3.9}$$

different ways to sample.

How many different ways can we sample r elements without replacement from a population with n elements?

3.5. Since ordering matters, we have

$$n(n-1)(\cdots)\big(n-(r-1)\big) = (n)_r \tag{3.10}$$

This defines the "**Falling Factorial**" $(n)_r$.

How can we be certain about this? Well, for the first choice there are n possibilities. The second has $(n-1)$ possibilities. The k^{th} has $(n-(k-1))$ possible choices. We then multiply these together to get the result.

3.6. Theorem. For a population with $n > 1$ elements and a prescribed sample size $r \leq n$, there exists n^r different samples with replacement, and $(n)_r$ samples without replacement.

3.7. Birthday Problem. How many people do we need in a room to make a favorable bet (i.e., a bet with probability of success greater than 1/2) that 2 people in the room will have the same birthday?

Solution: Let r be the number of people in the room, we want to find it. We will compute the probability of no repetition in our data (i.e., the probability no one in the room shares a birthday). The population, for this example, is the 365 days of the year.

Then

$$p_r = \frac{(365)_r}{365^r} \tag{3.11}$$

describes the probability that no one shares a birthday. We want to find the smallest r such that $p_r < 0.5$. This can be done by exhaustion, we find

$$\begin{aligned} p_{23} &= \frac{36997978566217959340182499134166757044383351847256064}{75091883268515350125426207425223147563269805908203125} \\ &\approx 0.49270276 \end{aligned} \tag{3.12}$$

So only make this bet when there are at least 23 people in the room!

3.3 Binomial Coefficients

3.8. How many samples of size r exist in a population of n elements, without regard to ordering? That is, how many ways can we choose r guys from n elements? We count this using binomial coefficients

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{3.13}$$

and read it as "n choose r". Why is this true? Great question: we have $(n)_r$ subpopulations, and $r!$ different orderings of each subpopulation, so we have

$$\binom{n}{r} = \frac{(n)_r}{r!}. \tag{3.14}$$

But this is precisely what we have written!

Convention. If $k \geq n$, then

$$\binom{n}{k} = 0. \tag{3.15}$$

3.9. Theorem. A population with n elements has $\binom{n}{r}$ different subpopulations of size $r \leq n$.

3.10. Theorem (Properties of the Binomial Coefficients). Let $0 \leq k \leq n$. Then

1. $\displaystyle\binom{n}{0} = \binom{n}{n} = 1$
2. $\displaystyle\binom{n}{k} = \binom{n}{n-k}$
3. Pascal's triangle $\displaystyle\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$

Proof. (1) We see by definition that

$$\binom{n}{0} = \frac{n!}{(n-0)!0!} = \frac{n!}{n!} = 1, \tag{3.16a}$$

and similarly

$$\binom{n}{n} = \frac{n!}{(n-n)!n!} = \frac{n!}{n!} = 1. \tag{3.16b}$$

That proves (1).

(2) We see that

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{3.17a}$$

and

$$\binom{n}{n-k} = \frac{n!}{\big(n-(n-k)\big)!(n-k)!} = \frac{n!}{k!(n-k)!} \tag{3.17b}$$

Then setting equals to equals proves (2).

(3) We recall the generating function for binomial coefficients is given by

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k, \tag{3.18a}$$

so

$$(1+x)^{n+1} = \sum_{k=0}^{n+1} \binom{n+1}{k} x^k. \tag{3.18b}$$

But simple arithmetic also suggests

$$\begin{aligned}(1+x)^n(1+x) &= 1 \cdot \sum_{k=0}^{n} \binom{n}{k} x^k + x \sum_{k=0}^{n} \binom{n}{k} x^k \\ &= \sum_{k=0}^{n} \binom{n}{k} x^k + \sum_{k=0}^{n} \binom{n}{k} x^{k+1}.\end{aligned} \tag{3.18c}$$

We can rewrite this as

$$\begin{aligned}(1+x)^n(1+x) &= 1 + \sum_{k=1}^{n} \binom{n}{k} x^k + \sum_{k=1}^{n} \binom{n}{k-1} x^k + x^{n+1} \\ &= 1 + \sum_{k=1}^{n} \left[\binom{n}{k} + \binom{n}{k-1}\right] x^k + x^{n+1}\end{aligned} \tag{3.18d}$$

and thus we conclude, setting Eq (3.18b) equal to Eq (3.18d), the coefficients of powers of x must be equal. Thus

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}, \tag{3.18e}$$

which implies (3). □

3.10.1. Corollary. *Let $n, k \in \mathbb{N}$, then*

$$\binom{n}{-k} = 0. \tag{3.19}$$

Proof of Corollary. By statement 2 of the previous theorem, we have

$$\binom{n}{-k} = \binom{n}{n+k} \tag{3.20}$$

Since $k > 0$, our convention implies that

$$\binom{n}{n+k} = 0 \tag{3.21}$$

as desired. □

▸ **Exercise 1.** Prove $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$.

3.11. Example. A deck of cards has 52 cards, a poker hand has 5 cards. How many different poker hands are there?

Solution: We see there are

$$\binom{52}{5} = 2\,598\,960 \tag{3.22}$$

different poker hands.

3.12. Example. The US Senate has 100 Senators (a pair represents each state). If a committee is formed with 50 senators, then:

(a) What's the probability a given state is represented?

(b) What's the probability each state is represented?

Solution. (a) Well, we first consider the *negation* of this statement: what's the probability not every state is represented? There are

$$\binom{98}{50} = d_k \tag{3.23}$$

ways to choose the committee so some state is not represented. But there is

$$\binom{100}{50} = n_k \tag{3.24}$$

different ways to choose the committee. Thus the probability a given state is not represented is

$$q = \frac{\binom{98}{50}}{\binom{100}{50}} = \frac{49}{198} \approx 0.24747\ldots \tag{3.25}$$

and hence the probability a given state is represented is

$$1 - q = \frac{149}{198} \approx 0.7525\ldots \tag{3.26}$$

That concludes the first question.

(b) We pick one senator from each state. We have two possibilities for each state, and 50 states, hence we have 2^{50} possibilities. Thus we have

$$p = \frac{2^{50}}{\binom{100}{50}} \approx 1.11 \times 10^{-14} \tag{3.27}$$

is the probability every state is represented.

3.3.1 Binomial Distribution

3.13. Example. Consider a random distribution of r balls in n urns. Find the probability p_k that a specified cell contains exactly k balls.

Solution: We want to find

$$p_k = \frac{\begin{pmatrix}\text{Number of ways to}\\ \text{choose } k \text{ balls from}\\ r \text{ balls}\end{pmatrix}\begin{pmatrix}\text{number of ways to}\\ \text{put remaining } (r-k) \text{ balls}\\ \text{in } (n-1) \text{ cells}\end{pmatrix}}{\begin{pmatrix}\text{number of different}\\ \text{ways placing } r\\ \text{balls in } n \text{ cells}\end{pmatrix}} \tag{3.28}$$

There are n^r different ways placing r balls in n cells. This gives us the denominator. We see that

$$\begin{pmatrix}\text{Number of ways to}\\ \text{choose } k \text{ balls from}\\ r \text{ balls}\end{pmatrix} = \binom{r}{k} \tag{3.29}$$

and what of the remaining $(r-k)$ balls? They can be placed in the remaining cells in $(n-1)^{r-k}$ ways. Thus we plug this all back into our equation to find

$$\begin{aligned} p_k &= \frac{\binom{r}{k}(n-1)^{r-k}}{n^r} \\ &= \binom{r}{k}\frac{1}{n^k}\left(1-\frac{1}{n}\right)^{r-k}. \end{aligned} \tag{3.30}$$

This is an example of the famous *binomial distribution*, which we'll soon study.

Binomial Distribution

3.14. Example. Suppose we toss a fair coin n times. What's the probability we have k heads?

Solution: The basic solution we will propose will look like

$$\Pr(k\text{ heads}) = \begin{pmatrix}\text{number of}\\ \text{combinations}\end{pmatrix}\begin{pmatrix}\text{probability of}\\ \text{one heads}\end{pmatrix}^{k}\begin{pmatrix}\text{probability of}\\ \text{one tails}\end{pmatrix}^{n-k} \tag{3.31}$$

We have a subpopulation of k heads in a population with n elements. So

$$\begin{pmatrix}\text{number of}\\ \text{combinations}\end{pmatrix} = \binom{n}{k} \tag{3.32}$$

and

$$\begin{pmatrix}\text{probability of}\\ \text{one heads}\end{pmatrix} = \begin{pmatrix}\text{probability of}\\ \text{one tails}\end{pmatrix} = \frac{1}{2}. \tag{3.33}$$

More generally, we will write

$$p = \begin{pmatrix}\text{probability of}\\ \text{one heads}\end{pmatrix},\quad\text{and}\quad (1-p) = \begin{pmatrix}\text{probability of}\\ \text{one tails}\end{pmatrix}. \tag{3.34}$$

So we conclude

$$\Pr(k\text{ heads}) = \binom{n}{k}p^k(1-p)^{n-k}. \tag{3.35}$$

Solution 2: Let p be the probability of success (heads), and q the probability of failure (tails), for one trial. Then we consider the Binomial expansion of

$$(p+q)^n = p^n + \cdots + \binom{n}{k}p^k q^{n-k} + \cdots + q^n. \tag{3.36}$$

Each term represents one outcome, namely, the coefficient of p^k is the outcome with k successes. So

$$\Pr(k\text{ heads}) = \binom{n}{k}p^k q^{n-k}. \tag{3.37}$$

Since the only outcomes are success or failure, we have

$$p + q = 1, \tag{3.38}$$

which implies $q = 1-p$ and "something must happen". Thus we conclude

$$\Pr(k\text{ heads}) = \binom{n}{k}p^k(1-p)^{n-k} \tag{3.39}$$

recovering our previous solution.

3.15. Example. In a ten-question true-false exam, find the probability that a student gets a grade of 70 percent or better by guessing. Answer the same question if the test has 30 questions.

Solution: When there are 10 questions, the student needs to answer 7 correctly. If the student guesses on a true-false exam, the student has a probability of success $p = 1/2$. Thus we see

$$\Pr(70\% \text{ with 10 questions}) = \binom{10}{7} 2^{-10} = \frac{15}{128}. \tag{3.40}$$

This is approximately $\Pr(70\% \text{ with 10q}) \approx 0.117$.

If there were 30 questions, we see

$$\begin{aligned}\Pr(70\% \text{ with 30 questions}) &= \binom{30}{21} 2^{-30} \\ &= \frac{3 \cdot 5^2 \cdot 11^2 \cdot 13 \cdot 23 \cdot 29}{2^{28}}\end{aligned} \tag{3.41}$$

which is approximately $\Pr(70\% \text{ with 30q}) \approx 0.293$.

3.4 Multinomial Coefficient

3.16. Suppose we have a population with n elements. How many different ways can we partition the population into subpopulations with k_1 members, ..., k_m members?

▸ **Exercise 2.** Prove that $k_1 + \cdots + k_m = n$.

We have the following number of such partitions:

$$\binom{n}{k_1, \ldots, k_m} = (k_1, \ldots, k_m)! = \frac{(k_1 + \cdots + k_m)!}{k_1!(\ldots)k_m!} \tag{3.42}$$

These quantities are called "**Multinomial Coefficients**". The notation varies according to the text.

3.5 Hat Check Problem

3.17. Problem Statement. At a restaurant, n people check their hats. The hat checker completely loses track of who owns which hat, and begins handing the hats randomly to the owners. What's the probability no one receives their hat back?

3.18. We really want to consider the permutations

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & \ldots & n \\ a_1 & a_2 & a_3 & \ldots & a_N \end{pmatrix} \tag{3.43}$$

such that there are *no fixed points* (i.e., $a_k \neq k$ for every k).

3.19. Solution. Let A_i be the event the i^{th} element is fixed. Then

$$A_1 \cup A_2 \cup \ldots A_n = \begin{pmatrix} \text{event there exists} \\ \text{at least one} \\ \text{fixed point} \end{pmatrix} \tag{3.44}$$

then

$$\Pr(\text{no fixed pt}) = 1 - \Pr(A_1 \cup A_2 \cup \ldots A_n). \tag{3.45}$$

We just have to compute $\Pr(A_1 \cup A_2 \cup \ldots A_n)$. What to do? Use inclusion-exclusion!

3.20. So we consider the event A_i. The permutation would look like

$$\begin{pmatrix}1 & 2 & 3 & \dots & i & \dots & n\\ & & & \dots & i & \dots & \end{pmatrix} \tag{3.46}$$

where the blank entries are arbitrary (we don't care where they go). How many different such permutations are there? Well, it's as though we permute the $(n-1)$ elements other than i, so there are $(n-1)!$ such permutations. The probability of this event is then

$$\Pr(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}. \tag{3.47}$$

There are n such events (since we can let $i = 1, \dots, n$). Thus

$$\sum_{i=1}^{n} \Pr(A_i) = n \cdot \frac{1}{n} = 1. \tag{3.48}$$

Wonderful.

3.21. Consider the event $A_i \cap A_j$, which has the permutation

$$\begin{pmatrix}1 & 2 & \dots & i & \dots & j & \dots & n\\ & & \dots & i & \dots & j & \dots & \end{pmatrix} \tag{3.49}$$

How many such permutations are there? Well, we fix two points while permuting the others, so there are $(n-2)!$ such permutations. Thus the probability is

$$\Pr(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)} \tag{3.50}$$

How many such "fix two point" events are there? It's simply

$$\binom{n}{2} = \frac{n(n-1)}{2} \tag{3.51}$$

thus

$$\sum_{1\leq i<j\leq n} \Pr(A_i \cap A_j) = \frac{n(n-1)}{2} \cdot \frac{1}{n(n-1)} = \frac{1}{2}. \tag{3.52}$$

This will be our second term.

3.22. Consider the event $A_i \cap A_j \cap A_k$. We see

$$\Pr(A_i \cap A_j \cap A_k) = \frac{(n-3)!}{n!} = \frac{1}{n(n-1)(n-2)}. \tag{3.53}$$

There are $\binom{n}{3}$ such terms. Thus

$$\sum_{1\leq i<j<k\leq n} \Pr(A_i \cap A_j \cap A_k) = \frac{1}{3!} \tag{3.54}$$

and inductively we find

$$\Pr(A_1 \cap \dots \cap A_n) = \frac{1}{n!} \tag{3.55}$$

Thus

$$\Pr(\text{no fixed pt}) = \frac{1}{2} - \frac{1}{3!} + \frac{1}{4!} + \dots + \frac{(-1)^n}{n!} \tag{3.56}$$

We should note the Taylor series for e^x is

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + \dots \tag{3.57}$$

so

$$\mathrm{e}^{-1} = 1 - 1 + \frac{(-1)^2}{2!} + \dots + \frac{(-1)^n}{n!} + \dots \tag{3.58}$$

implies

$$\Pr(\text{no fixed pt}) \to \mathrm{e}^{-1} \tag{3.59}$$

as $n \to \infty$.

We will construct a table, let p_n be the probability there are no fixed points among permutations of n elements. Then:

$$\begin{aligned}
p_2 &= \frac{1}{2} \approx 0.5\\
p_3 &= \frac{1}{3} \approx 0.33333334\\
p_4 &= \frac{3}{8} \approx 0.375\\
p_5 &= \frac{11}{30} \approx 0.36666667\\
p_6 &= \frac{53}{144} \approx 0.36805555\\
p_7 &= \frac{103}{280} \approx 0.36785713\\
p_8 &= \frac{2119}{5760} \approx 0.36788195\\
p_9 &= \frac{16687}{45360} \approx 0.36787918
\end{aligned} \tag{3.60}$$

We should note that $1/\mathrm{e} \approx 0.36787944117144233$, so we are getting a decent approximation with $n = 8$ (five digits!).

3.6 Hypergeometric Distribution

3.23. The "**Hypergeometric Distribution**" occurs in situations like:

> Suppose we have N balls, of which k are red and $N - k$ are blue. We draw a sample, without replacement, of n balls. Let X be the number of red balls drawn in our sample of size n. What's the probability $X = x$?

We see

$$\Pr(X = x) = \frac{\begin{pmatrix}\text{number of different}\\ \text{ways to choose}\\ x\text{ red balls}\end{pmatrix}\begin{pmatrix}\text{number of different}\\ \text{ways to choose}\\ (n-x)\text{ blue balls}\end{pmatrix}}{\begin{pmatrix}\text{number of different}\\ \text{samples drawn}\end{pmatrix}} \tag{3.61}$$

3.24. Example. We have 1000 widgets, of which an unknown number D has defects. A sample of 100 has 2 with defects. The "**Maximum Likelihood Estimate**" for D is the number which gives the highest probability for obtaining the number of defectives observed in a sample. Find that value of D.

Solution: So we have $N = 1000$ and instead of "red balls" we have "defective Widgets" $k = D$. The sample size is $n = 100$. What's the value of D that makes the event most probable?

Well, the distribution would be described by

$$\Pr(X = 2) = \frac{\binom{D}{2}\binom{1000-D}{100-2}}{\binom{1000}{100}} \tag{3.62}$$

which algebraically reduces to

$$\Pr(X = 2) = \frac{D(D-1)}{2}\frac{1}{100\cdot 999\cdot(\dots)\cdot(1000-(D-1))}\frac{900\cdot(\dots)\cdot(902-(D-1))}{1}\cdot 100\cdot 99. \tag{3.63}$$

We can then write up a small C program which will write out a table for values of D and the corresponding probability.

```c
#include <stdlib.h>
#include <stdio.h>
#include <math.h>

#define E 2.71828182845904523536028747135266249775724709369995L
#define PI 3.14159265358979323846264338327950288L

/* Stirling's approximation */
double factorial(int n)
{
  double x = 1.0L*n;
  return sqrt(2*PI*x)*pow(x/E,x);
}

/* return a*(a+1)*(...)*b */
double prod(int a, int b)
{
  if(a>b) return prod(b,a);
  int k;
  double result;
  result = 1.0;
  for(k=a;k<=b;k++)
    result = result*k;
  return result;
}

int main(int argc, char *argv[])
{
  int D;
  double c, v;
  v = 0.0;
  c = 990.0;
  for(D=2; 35>D; D++)
  {
    v = c * (D*(D-1)*0.5) * prod(902-D+1,900)/prod(1000-D+1,1000);
    printf("D=%d,␣Pr(X=2)␣=␣%f\n",D,v);
  }

  return EXIT_SUCCESS;
}
```

A small snippet reveals:

```
D=19, Pr(X=2) = 0.028804
D=20, Pr(X=2) = 0.028807
D=21, Pr(X=2) = 0.028655
D=22, Pr(X=2) = 0.028366
D=23, Pr(X=2) = 0.027954
```

which implies $D = 20$ is the value which has the most probable outcome.

3.25. Example. On an Island, 50 moose are captured and tagged. Six months later, 200 moose are captured, of which 8 are tagged. Estimate the number of moose on the Island.

Solution: So, this is a hypergeometric distribution, where the "red balls" are the tagged moose, we have a sample of 200 without replacement, and 8 of them are tagged. So we are trying to maximize the function

$$h(N, 50, 200, 8) = \frac{\binom{50}{8}\binom{N-50}{200-8}}{\binom{N}{200}} \tag{3.64}$$

by picking some N. After some algebra, we can rewrite this as

$$\begin{aligned} h(N, 50, 200, 8) &= \frac{50!}{42!}\frac{1}{8!}\frac{200!}{192!}\frac{(N-50)!}{N!}\frac{(N-200)!}{(N-242)!} \\ &= C\frac{(N-50)!}{N!}\frac{(N-200)!}{(N-242)!} \end{aligned} \tag{3.65}$$

where $C = 1192725059258223539848800000$. We have $N \geq 242$. We write up a small program, listed below, to write out a table of probabilities. It's maximized when $N = 1250$.

```
#include <stdlib.h>
#include <stdio.h>
#include <math.h>

/* return a*(a+1)*(...)*b */
long double prod(int a, int b)
{
  if(a>b) return prod(b,a);
  int k;
  long double result;
  result = 1.0;
  for(k=a;k<=b;k++)
    result = result*k;
  return result;
}

int main(int argc, char *argv[])
{
  int N;
  long double c, v;
  v = 0.0;
  c = 1192725059258223539848800000.0L;
  for(N=1242; 1260>N; N++)
  {
    v = c * prod(N-200,N-241)/prod(N-49,N);
    printf("N=%d,␣h(N,50,200,8)␣=␣%Lf\n",N,v);
  }

  return EXIT_SUCCESS;
}
```

3.26. Example. Suppose that in a bushel of 550 apples there are 2% rotten ones. What is the probability that a random sample of 25 apples contains two rotten apples? Hint: Hypergeometric distribution.

Solution: So, we have $N = 550$ apples, of which $k = 11$ are rotten. So we pick a sample $n = 25$. What's the probability $x = 2$ are rotten? It's given by

$$\Pr(2\text{ rotten}) = \frac{\binom{11}{2}\binom{539}{23}}{\binom{550}{25}} \tag{3.66}$$

We can compute this by hand, finding

$$\begin{aligned}\Pr(2\text{ rotten}) &= \frac{599494391824595575}{8092091399320955412}\\ &\approx 0.074083986\end{aligned} \tag{3.67}$$

so the probability is roughly 7.4%.

4 Partial Summation: A Useful Tool

4.1. Definition. Let $x \in \mathbb{R}$ be any real number. We define $[x]$ to be the greatest integer smaller than x. So observe

$$[3.1] = 3 \tag{4.1a}$$

$$[\pi] = 3 \tag{4.1b}$$

$$[-\mathrm{e}] = -3 \tag{4.1c}$$

We always will have

$$[x] \leq x. \tag{4.2}$$

We now may define

$$\{x\} = x - [x] = \text{fractional part of } x \tag{4.3}$$

and observe $0 \leq \{x\} < 1$.

4.2. Lemma. *Suppose we have a sequence $\{c_n\}_{n=1}^{\infty}$ and for $x \geq 1$ we define a function*

$$C(x) = \sum_{n\leq x} c_n, \quad \text{and} \quad C(0) = 0. \tag{4.4}$$

Let f be a C^1 function, then

$$\sum_{n\leq x} c_n f(n) = C(x)f(x) - \int_1^x C(t)f'(t)\,\mathrm{d}t. \tag{4.5}$$

Proof. This is a two-step proof. Step one notes

$$\sum_{n\leq x} C(n)\big(f(n+1) - f(n)\big) = C(x)f([x]) - \sum_{n\leq x} c_n f(n) \tag{4.6a}$$

but the left hand side is precisely

$$\sum_{n\leq x} C(n)\big(f(n+1) - f(n)\big) = \int_1^{[x]} C(t)f'(t)\,\mathrm{d}t. \tag{4.6b}$$

Thus we obtain

$$\int_1^{[x]} C(t)f'(t)\,\mathrm{d}t = C(x)f([x]) - \sum_{n\leq x} c_n f(n). \tag{4.6c}$$

That concludes the first step.

The second step notes

$$\int_{[x]}^{x} C(t)f'(t)\,\mathrm{d}t = C(x)f(x) - C(x)f([x]). \tag{4.7}$$

This uncontroversial statement should be seen immediately by the fundamental theorem of calculus.

We then add Eq (4.6c) to Eq (4.7) to find

$$\begin{aligned}\int_{[x]}^{x} C(t)f'(t)\,\mathrm{d}t + \int_{1}^{[x]} C(t)f'(t)\,\mathrm{d}t&\\ = C(x)f(x) - C(x)f([x]) + C(x)f([x]) - \sum_{n\leq x} c_n f(n)&\end{aligned} \tag{4.8a}$$

which simplifies to

$$\int_{1}^{x} C(t)f'(t)\,\mathrm{d}t = C(x)f(x) + \sum_{n\leq x} c_n f(n) \tag{4.8b}$$

precisely as desired. □

4.3. Example (Harmonic Series). We will apply our lemma to the harmonic series. How? Well, consider

$$H_x = \sum_{n\leq x}\frac{1}{n} \tag{4.9}$$

which we consider the sequence $c_n = 1$ and $f(t) = 1/t$. Then note

$$C(x) = \sum_{n\leq x} 1 = [x]\,. \tag{4.10}$$

Thus our lemma implies

$$H_x = \frac{[x]}{x} + \int_{1}^{x}\frac{[t]}{t^2}\,\mathrm{d}t. \tag{4.11}$$

We will try to simplify this.

First we should note that $x = [x] + \{x\}$. Thus

$$\begin{aligned}H_x &= \frac{x-\{x\}}{x} + \int_{1}^{x}\frac{t-\{t\}}{t^2}\mathrm{d}t &(4.12\text{a})\\ &= \ln(x) + \left(1 - \frac{\{x\}}{x}\right) - \int_{1}^{x}\frac{\{t\}}{t^2}\mathrm{d}t &(4.12\text{b})\\ &= \ln(x) + \left(1 - \int_{1}^{\infty}\frac{\{t\}}{t^2}\mathrm{d}t\right) + \int_{x}^{\infty}\frac{\{t\}}{t^2}\mathrm{d}t - \frac{\{x\}}{x} &(4.12\text{c})\end{aligned}$$

Let

$$\gamma = 1 - \int_{1}^{\infty}\frac{\{t\}}{t^2}\mathrm{d}t \tag{4.13}$$

then our expression simplifies to

$$H_x = \ln(x) + \gamma + \int_{x}^{\infty}\frac{\{t\}}{t^2}\mathrm{d}t - \frac{\{x\}}{x}. \tag{4.14}$$

Wonderful.

But does the integral term converge? We see

$$\left|\int_x^\infty \frac{\{t\}}{t^2}\mathrm{d}t\right| \leq \int_x^\infty \frac{|\{t\}|}{t^2}\mathrm{d}t \tag{4.15a}$$

$$\leq \int_x^\infty \frac{1}{t^2}\mathrm{d}t = \frac{1}{x}. \tag{4.15b}$$

Moreover this implies

$$\int_x^\infty \frac{\{t\}}{t^2}\mathrm{d}t - \frac{\{x\}}{x} = \mathcal{O}(x^{-1}). \tag{4.16}$$

So the expression for H_x becomes

$$H_x = \ln(x) + \gamma + \mathcal{O}(x^{-1}) \tag{4.17}$$

but *can we do better?*

4.4. Lets try to figure out the next terms to order $\mathcal{O}(x^{-2})$. We will consider

$$\int_x^\infty \frac{\{t\}}{t^2}\mathrm{d}t - \frac{\{x\}}{x} = \int_x^\infty \frac{\{t\}-\frac{1}{2}}{t^2}\mathrm{d}t + \frac{1}{2}\int_x^\infty \frac{\mathrm{d}t}{t^2} - \frac{\{x\}}{x} \tag{4.18}$$

which simplifies to

$$\int_x^\infty \frac{\{t\}}{t^2}\mathrm{d}t - \frac{\{x\}}{x} = \int_x^\infty \frac{\{t\}-\frac{1}{2}}{t^2}\mathrm{d}t + \frac{1}{2x} - \frac{\{x\}}{x}. \tag{4.19}$$

Now we make a claim!

Claim. The integral $\displaystyle\int_x^\infty \frac{\{t\}-\frac{1}{2}}{t^2}\mathrm{d}t = \mathcal{O}(1/x^2)$.

Proof. We will integrate by parts, using

$$u = \frac{1}{t^2}, \quad \text{and} \quad \mathrm{d}u = \frac{-2}{t^3}\mathrm{d}t \tag{4.20}$$

and

$$\mathrm{d}v = \left(\frac{1}{2} - \{t\}\right)\mathrm{d}t, \quad \text{and} \quad v = \int_1^t \left(\frac{1}{2} - \{u\}\right)\mathrm{d}u. \tag{4.21}$$

Thus the integral becomes

$$\lim_{R\to\infty}\int_x^R \frac{\{t\}-\frac{1}{2}}{t^2}\mathrm{d}t$$

$$= \frac{1}{t^2}\int_1^t \left(\frac{1}{2} - \{u\}\right)\mathrm{d}u\Bigg|_{t=x}^{t=R} + 2\int_x^R \frac{1}{t^3}\int_1^t \left(\frac{1}{2} - \{u\}\right)\mathrm{d}u \tag{4.22}$$

We claim that

$$\begin{aligned} v &= \int_1^{[t]} \left(\frac{1}{2} - \{u\}\right)\mathrm{d}u + \int_{[t]}^t \left(\frac{1}{2} - \{u\}\right)\mathrm{d}u \\ &= 0 + \int_{[t]}^t \left(\frac{1}{2} - \{u\}\right)\mathrm{d}u \end{aligned} \tag{4.23}$$

How can we see this? Well, we should note for any integer k that

$$\int_k^{k+1} \left(\frac{1}{2} - \{u\}\right)\mathrm{d}u = 0, \tag{4.24}$$

since it's a sawtooth function. Thus

$$v \leq \int_{1/2}^{1} \left(\frac{1}{2} - \{u\}\right) \mathrm{d}u = \frac{1}{4}. \tag{4.25}$$

Our integral becomes

$$\begin{aligned}\lim_{R\to\infty} \int_{x}^{R} \frac{\{t\} - \frac{1}{2}}{t^2} \mathrm{d}t = \lim_{R\to\infty} \frac{1}{t^2}\frac{1}{4}\Big|_{t=x}^{t=R} + 2\int_{x}^{R} \frac{1}{t^3}\frac{1}{4}\mathrm{d}t \\ \sim \mathcal{O}(x^{-2}).\end{aligned} \tag{4.26}$$

This proves the claim. □

Moreover, for any integer n, we see

$$\sum_{n=1}^{N} \frac{1}{n} = \ln(N) + \gamma + \frac{1}{2N} + \mathcal{O}(N^{-2}). \tag{4.27}$$

This turns out to be a good *asymptotic* approximation for harmonic numbers.

5 Random Variables

5.1. In an election, 50 people vote. We describe all the outcomes in a sample space Ω, but don't we have a valid question "How many voted 'yes'?"?

How can we answer such a question? *Count the number of 'yes'-es!* This is done with a map

$$Y \colon \Omega \to \mathbb{N}_0. \tag{5.1}$$

But what is this mapping? The number of 'yes'-es, which depends on the event. It's a *random variable!*

Note that we can extend the codomain from $\mathbb{N}_0$ to $\mathbb{Z}$, $\mathbb{Q}$, or $\mathbb{R}$. For the sake of generality[5], we will consider $\mathbb{R}$.

5.2. Definition. Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. We define a "**Random Variable**" to be a function $W \colon \Omega \to \mathbb{R}$ such that for each $x \in \mathbb{R}$ we have the set

$$X = \{\omega \in \Omega : W(\omega) \leq x\} \tag{5.2}$$

be an element of $\mathcal{F}$, or in symbols $X \in \mathcal{F}$.

5.2.1. Remark. We will consider the simpler case of "**Discrete Random Variables**" $X \colon \Omega \to \mathbb{Z}$.

The really critical theoretic property for random variables X is for any interval $B \subseteq \mathbb{R}$, we have a set of outcomes

$$\{\omega \in \Omega : X(\omega) \in B\} \tag{5.3}$$

(denoted $X \in B$) which lives in the σ-algebra. Studying $\Pr(X \in B)$ will become increasingly relevant.

5.3. Examples. Lets give a grocery list of examples.

1. Toss a coin N times, let H be the number of heads.
2. Choose a random point on $\mathbb{R}^n$, let D be the distance from the point to the origin.

[5] We can sometimes include $+\infty$, or $-\infty$, if necessary.

3. Take a random person from a class, let X be the student's height.

4. Let W be the value of the DOW stock index at closing.

5.4. A discrete random variable has countably many values $\{x_i \in \mathbb{R} : i \in I \subseteq \mathbb{Z}\}$. We take the convention that its codomain is a subdomain of $\mathbb{Z}$. Let X be a discrete random variable, then its "**Probability Mass Function**" $p(x_i) := \Pr(X = x_i)$.

5.5. Proposition. A probability mass function p satisfies:

1. For any i, $p(x_i) > 0$

2. For any interval B, $\Pr(X \in B) = \sum_{x_i \in B} p(x_i)$

3. We have $\sum_i p(x_i) = 1$.

5.6. Example. Let X be the number of heads in 2 fair coin tosses. What is its probability mass function?

Solution: We see there are three outcomes: 0, 1, 2. We also see that $\Pr(X = 0) = 1/4$ and $\Pr(X = 2) = 1/4$. Thus $\Pr(X = 1) = 1/2$. This gives us the probability mass function.

5.7. Example. An urn contains 20 slips of paper numbered 1, …, 20. We select 5 at random, without replacement. Let X be the random variable describing the greatest value of the 5 slips selected.

(a) Determine the probability mass function for X.

(b) What's the probability at least one of the slips selected is 15 or greater?

Solution. Well, X takes the values 5, …, 20. There are $\binom{20}{5}$ different outcomes. So we see

$$\Pr(X \leq k) = \frac{\binom{k}{5}}{\binom{20}{5}}. \tag{5.4}$$

Thus

$$p(k) = \Pr(X \leq k) - \Pr(X \leq k - 1) \tag{5.5}$$

and using Pascal's triangle (theorem 3.10), we have

$$\begin{aligned} p(k) &= \frac{\binom{k}{5} - \binom{k-1}{5}}{\binom{20}{5}} \\ &= \frac{\binom{k-1}{4}}{\binom{20}{5}} \end{aligned} \tag{5.6}$$

This gives us the probability mass function.

(b) The probability one of the slips is 15 or greater can be calculated using

$$\Pr(X \geq 15) = \sum_{k=15}^{20} p(k). \tag{5.7}$$

Equivalently, we can calculate it as

$$\begin{aligned} \Pr(X \geq 15) = 1 - \Pr(X \leq 14) &= \frac{715}{15504} \\ &\approx 0.046117 \end{aligned} \tag{5.8}$$

or less than a 1/20 probability.

5.8. Definition. Let X be a discrete random variable taking values x_1, x_2, ... Then the "**Expected Value**" (also called the *average* or *mean* or *expectation*) for X is

$$\mathrm{E}[X] = \sum_i x_i \Pr(X = x_i). \tag{5.9}$$

We also have, for any function $g\colon \mathbb{R} \to \mathbb{R}$,

$$\mathrm{E}[g] = \sum_i g(x_i) \Pr(X = x_i). \tag{5.10}$$

5.9. Example. Let X be a random variable such that $\Pr(X = 1) = 0.2$, $\Pr(X = 2) = 0.3$ and $\Pr(X = 3) = 0.5$. What's the expected value of X?

Solution: We find, using our definition,

$$\mathrm{E}[X] = 1 \Pr(X = 1) + 2 \Pr(X = 2) + 3 \Pr(X = 3) \tag{5.11}$$

and this becomes

$$\begin{aligned} \mathrm{E}[X] &= 1 \cdot 0.2 + 2 \cdot 0.3 + 3 \cdot 0.5 \\ &= 0.2 + 0.6 + 1.5 = 2.3. \end{aligned} \tag{5.12}$$

That concludes our example.

5.10. Given some discrete random variable X and its expected value $\mu = \mathrm{E}[X]$, how can we measure the "spread" of (the probability mass function for) X? The naïve solution would use $\mathrm{E}|X - \mu|$, but this is bad since we should avoid absolute values.

So we define the "**Variance**" of X as

$$\mathrm{Var}(X) = \mathrm{E}(X - \mu)^2. \tag{5.13}$$

Notice this intuitively measures the sum of the "distance squared" of the values x_i from the expected value μ. This has the "wrong units" (distance squared as opposed to distance). So we just take the square root, and we obtain the "**Standard Deviation**"

$$\sigma(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\mathrm{E}(X - \mu)^2}. \tag{5.14}$$

5.11. Proposition. The expected value is linear, i.e., given random variables X and Y and any coefficients c_1, c_2 we have

$$\mathrm{E}[c_1 X + c_2 Y] = c_1 \mathrm{E}[X] + c_2 \mathrm{E}[Y]. \tag{5.15}$$

5.12. Also note that

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[(X - \mu)^2] && (5.16a) \\ &= \mathrm{E}[X^2 - 2\mu X + \mu^2] && (5.16b) \\ &= \mathrm{E}[X^2] - 2\mu \mathrm{E}[X] + \mu^2 && (5.16c) \\ &= \mathrm{E}[X^2] - 2\mu^2 + \mu^2 = \mathrm{E}[X^2] - (\mathrm{E}[X])^2. && (5.16d) \end{aligned}$$

This gives us another intuition for variance!

5.13. A “**Uniform Discrete Random Variable**” is a random variable taking values x_1, x_2, …, x_n each with equal probability $1/n$. Such a random variable simply takes a random choice among n numbers. Note that

$$\mathrm{E}[X] = \frac{x_1 + \cdots + x_n}{n} \tag{5.17a}$$

and

$$\mathrm{Var}[X] = \frac{x_1^2 + \cdots + x_n^2}{n} - \left(\frac{x_1 + \cdots + x_n}{n}\right)^2. \tag{5.17b}$$

5.14. Example. Let X be the number shown on a rolled fair die. What's $\mathrm{E}[X]$ and $\mathrm{Var}(X)$?

Solution: We find

$$\begin{aligned}\mathrm{E}[X] &= \frac{1 + 2 + \cdots + 6}{6} \\ &= \frac{1}{6}\left(\frac{6(7)}{2}\right) = \frac{7}{2},\end{aligned} \tag{5.18}$$

and

$$\mathrm{E}[X^2] = \frac{1 + 2^2 + \cdots + 6^2}{6} = \frac{91}{6} \tag{5.19}$$

which then implies

$$\begin{aligned}\mathrm{Var}(X) &= \frac{91}{6} - \left(\frac{7}{2}\right)^2 \\ &= \frac{91}{6} - \frac{49}{4} = \frac{70}{24} = \frac{35}{12}\end{aligned} \tag{5.20}$$

5.1 Bernoull Random Variable

5.15. Let A be an event with probability $p = \Pr(A)$. We have an “**Indicator Function**” for A be a function defined as

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases} \tag{5.21}$$

Note I_A is a random variable, since it's a function from $\Omega \to \mathbb{R}$.

Claim: $\mathrm{E}[I_A] = p$. Really? Well, observe

$$\begin{aligned}\mathrm{E}[I_A] &= I_A(A)\Pr(A) + I_A(A^{\mathsf{C}})\Pr(A^{\mathsf{C}}) \\ &= 1 \cdot p + 0 \cdot (1 - p) = p.\end{aligned} \tag{5.22}$$

What is its variance? We see

$$\mathrm{E}[I_A^2] = 1^2 \Pr(A) + 0^2 \Pr(A^{\mathsf{C}}) = p, \tag{5.23}$$

thus

$$\mathrm{Var}(I_A) = p - p^2 = p(1 - p). \tag{5.24}$$

5.16. Now, recall the “inclusion-exclusion” property for probability suggests

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \tag{5.25}$$

which can be derived using expectation values of Bernoulli random variables. Recall

$$I_A + I_{A^\complement} = 1 \tag{5.26}$$

where we abuse notation and write 1 for the constant function I_Ω. Now, we also see

$$I_{A\cap B}(x) = \begin{cases} 1 & \text{if } x\in A \text{ and } x\in B \\ 0 & \text{otherwise} \end{cases} \tag{5.27}$$

thus

$$I_{A\cap B} = I_A I_B. \tag{5.28}$$

Great.

We now want to consider $I_{A\cup B}$ in terms of I_A, I_B, and $I_{A\cap B}$. Observe

$$I_{A\cup B} = 1 - I_{(A\cup B)^\complement} = 1 - I_{A^\complement\cap B^\complement} \tag{5.29}$$

replacing $I_{A^\complement\cap B^\complement}$ with the product of indicator functions gives us

$$I_{A\cup B} = 1 - I_{A^\complement} I_{B^\complement}. \tag{5.30}$$

Using the fact $I_{A^\complement} = 1 - I_A$, we have

$$I_{A\cup B} = 1 - (1 - I_A)(1 - I_B). \tag{5.31}$$

Using basic algebra, expanding out the right hand side, we find

$$I_{A\cup B} = I_A + I_B - I_{A\cap B}. \tag{5.32}$$

Now, we take expectation values to find

$$\begin{aligned}\Pr(A\cup B) = \mathrm{E}(I_{A\cup B}) &= \mathrm{E}[I_A] + \mathrm{E}[I_B] - \mathrm{E}[I_{A\cap B}] \\ &= \Pr(A) + \Pr(B) - \Pr(A\cap B).\end{aligned} \tag{5.33}$$

This gives us an alternate proof of the inclusion-exclusion principle.

5.17. Hat Checker Problem Redux. We can use indicator functions to solve the hat checker problem, which we introduced in subsection 3.5. Let π be a permutation of n elements, we want to find the number of fixed points. Let

$$X(\sigma) = \text{number of fixed points of } \sigma. \tag{5.34}$$

We introduce the indicator function

$$I_i(\pi) = \begin{cases} 1 & \text{if } i \text{ is a fixed point of } \pi \\ 0 & \text{otherwise} \end{cases} \tag{5.35}$$

We now fix a number r such that $0 \leq r \leq n-2$. Why $n-2$? We want to sum over the permutations which do not fix all the points. If we fix $n-1$ points, then we fix all the points (think about it: if we give $n-1$ people their hats correctly, the remaining hat must belong to the remaining person!).

So we now set

$$\mathcal{S}_r = \sum_{\pi\in S_n} I_{j_1}(\dots)I_{j_r}(1 - I_{k_{r+1}})(\dots)(1 - I_{k_n}) \tag{5.36}$$

where we sum over all permutations. Observe there are $n!$ terms in the sum.

If $X(\sigma) \neq r$, we claim $\mathcal{S}_r = 0$. Why? Well, if $X(\sigma) > r$, then one of the $(1 - I_{k_*})$ factors vanishes in every term. If $X(\sigma) < r$, then one of the I_{j_*} factors vanishes in every term.

How many different scenarios do we have $X(\sigma) = r$? There are $r!$ different ways to have fixed points, and $(n-r)!$ different ways to permute the non-fixed points. Thus

$$\mathcal{S}_r = r!(n-r)! \tag{5.37}$$

as desired. So

$$\mathcal{S}_r(\sigma) = \begin{cases} r!(n-r)! & \text{if } X(\sigma) = r \\ 0 & \text{otherwise} \end{cases} \tag{5.38}$$

Thus we can construct an indicator function

$$\mathcal{I}_r = \frac{\mathcal{S}_r}{r!(n-r)!} \tag{5.39}$$

which tells us if a permutation has r fixed points.

Puzzle: What's $\mathrm{E}[\mathcal{I}_r]$?

Using linearity, we have

$$\mathrm{E}[\mathcal{I}_r] = \frac{1}{r!(n-r)!}\mathrm{E}[\mathcal{S}_r], \tag{5.40}$$

and we need to compute $\mathrm{E}[\mathcal{S}_r]$. We find

$$\mathrm{E}[\mathcal{S}_r] = \mathrm{E}\left(\sum_{\pi} I_{j_1}(\dots)I_{j_r}(1 - I_{k_{r+1}})(\dots)(1 - I_{k_n})\right) \tag{5.41}$$

then we expand the $(1 - I_{k_*})$ factors

$$\begin{aligned}\mathrm{E}[\mathcal{S}_r] = \mathrm{E}\Bigg(\sum_{\pi} I_{j_1}(\dots)I_{j_r} \times \Big[1 - (I_{k_{r+1}}I_{k_{r+2}} + \cdots + I_{k_{n-1}}I_{k_n})\\ + (\dots) + (-1)^{n-r}I_{k_{r+1}}(\dots)I_{k_n}\Big]\Bigg)\end{aligned} \tag{5.42}$$

Now, we claim that

$$\mathrm{E}[I_{j_1}(\dots)I_{j_r}I_{k_{r+s}}] = \mathrm{E}[I_{j_1}(\dots)I_{j_r}I_{k_{r+1}}] \tag{5.43}$$

for $s = 1, \dots, n-r$. This means our sum becomes

$$\mathrm{E}[\mathcal{S}_r] = \sum_{\pi}\sum_{s=0}^{n-r}(-1)^s\binom{n-r}{s}\mathrm{E}[I_{j_1}(\dots)I_{j_r}I_{k_{r+1}}(\dots)I_{k_{r+s}}] \tag{5.44}$$

where $0 \leq s \leq n-r$. So we see

$$\mathrm{E}[I_{j_1}(\dots)I_{j_r}I_{k_{r+1}}(\dots)I_{k_{r+s}}] = \frac{(n-r-s)!}{n!} \tag{5.45}$$

as there are $n!$ permutations with only $(n-r-s)!$ permutations fixing the given entries.

Now, we combine everything together, and find

$$\begin{aligned}
\mathrm{E}[\mathcal{I}_r] &= \frac{1}{r!(n-r)!}\mathrm{E}[\mathcal{S}_r] && (5.46a)\\
&= \frac{1}{r!(n-r)!}\sum_{\pi}\sum_{s=0}^{n-r}(-1)^s\binom{n-r}{s}\frac{(n-r-s)!}{n!} && (5.46b)\\
&= \frac{1}{r!(n-r)!}n!\sum_{s=0}^{n-r}(-1)^s\binom{n-r}{s}\frac{(n-r-s)!}{n!} && (5.46c)\\
&= \frac{1}{r!}\sum_{s=0}^{n-r}\frac{(-1)^s}{s!} && (5.46d)
\end{aligned}$$

So, summarizing the main conclusion, the probability that a random permutations has exactly r fixed points is given by

$$\Pr(X=r) = \frac{1}{r!}\left(\frac{1}{2!}-\frac{1}{3!}+\cdots+\frac{(-1)^{n-r}}{(n-r)!}\right). \tag{5.47}$$

This holds for $0 \leq r \leq n-2$. For $r=0$, this converges quickly to the value

$$\frac{1}{\mathrm{e}} = 0.367879441171\dots \tag{5.48}$$

We have thus derived another solution to the hat check problem using Bernoulli random variables.

5.2 Poisson Distribution

5.18. Derivation. Consider the binomial distribution

$$b(n,p) = \binom{n}{k}p^k(1-p)^{n-k}. \tag{5.49}$$

We take $\lambda = np$ to be the "intensity". Then we can re-write the distribution as

$$\begin{aligned}
b(n,p) &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} && (5.50a)\\
&= \left[\frac{n-(k-1)}{n}(\cdots)\frac{n-(k-k)}{n}\right]\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left[\left(1-\frac{\lambda}{n}\right)^{-k}\right] && (5.50b)\\
&= \left[\left(1-\frac{(k-1)}{n}\right)(\cdots)\left(1-\frac{(k-k)}{n}\right)\right]\\
&\qquad\times\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left[\left(1-\frac{\lambda}{n}\right)^{-k}\right]. && (5.50c)
\end{aligned}$$

Now we take the limit $n \to \infty$ while fixing k:

$$\begin{aligned}
\lim_{n\to\infty} b(n,p) &= \lim_{n\to\infty}\left[\left(1-\frac{(k-1)}{n}\right)(\cdots)\left(1-\frac{(k-k)}{n}\right)\right]\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left[\left(1-\frac{\lambda}{n}\right)^{-k}\right]\\
&= [1]\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}[1]. && (5.51)
\end{aligned}$$

This gives us the "**Poisson distribution**" *Poisson distribution*

$$f(k,\lambda) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}. \tag{5.52}$$

▸ **Exercise 3.** Consider the sample space $\Omega = \mathbb{N}_0$ obtained by the Poisson process. Calculate $\Pr(\Omega)$.

Solution: We see that

$$\Pr(\Omega) = \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} \tag{5.53}$$

but we can rearrange factors, and find

$$\Pr(\Omega) = \mathrm{e}^{-\lambda}\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = 1. \tag{5.54}$$

Thus, in the $f(k,\lambda)$, the factor $\mathrm{e}^{-\lambda}$ is a normalization constant.

5.19. Interpretation. We interpret λ as the number of events in some unit of time. The parameter k indicates the number of events we wonder about (i.e., we ask "What's the probability k events will happen?").

5.20. Example. Airlines find that passangers who make reservations fail to appear with probability 1/10, independent of other passangers. Acme Airlines sell 10 tickets for their 9 seat airplane, and 20 tickets for their 18 seat plane. Which plane is often over-booked?

Solution One: Lets write

$$f(n,k) = \binom{n}{k}\left(\frac{1}{10}\right)^k\left(1-\frac{1}{10}\right)^{n-k} \tag{5.55}$$

for the probability k people with reservations don't appear. We see the probability the smaller plane is overbooked occurs only when all 10 people appear, which has probability

$$f(10,0) = \left(\frac{9}{10}\right)^{10} = 0.3486784401. \tag{5.56}$$

The larger plane is overbooked when 19 or 20 people appear, which occurs with probability

$$\begin{aligned} f(20,0)+f(20,1) &= 20\frac{1}{10}\left(\frac{9}{10}\right)^{19} + \left(\frac{9}{10}\right)^{20} \\ &= 0.39174699812516770581. \end{aligned} \tag{5.57}$$

We see the larger plane gets overbooked a tad more often than the smaller plane.

Solution Two: Using the Poisson distribution, we will count the number of passangers that are absent as the "intensity". So $\lambda_s = n\cdot 1/10$ or $\lambda_s = 1$. We see

$$\mathrm{Poi}(k,\lambda_s) = \frac{\mathrm{e}^{-1}}{k!} \tag{5.58}$$

and we get overbooked for

$$\mathrm{Poi}(0,\lambda_s) = \mathrm{e}^{-1} \approx 0.36787944 \tag{5.59}$$

which is within 2% of Eq (5.56).

But for the larger plane, we see $\lambda_l = 2$ is the intensity of absent people, so

$$\mathrm{Poi}(0,\lambda_l) = \mathrm{e}^{-2} \tag{5.60a}$$

and

$$\mathrm{Poi}(1,\lambda_l) = 2\mathrm{e}^{-2} \tag{5.60b}$$

thus

$$\mathrm{Poi}(19,\lambda_l) + \mathrm{Poi}(20,\lambda_l) = 3\mathrm{e}^{-2} \approx 0.4060058 \tag{5.60c}$$

which is within 1.5% of Eq (5.57). We see these approximations are quite good!

5.3 Expected Values

5.21. If we have a discrete random variable $X\colon\Omega\to\mathbb{R}$, we can ask what's its *expected value*? What does this mean? We mean, if $X = x_j$ for $j\in\mathbb{N}$, we want to consider the expression

$$\mathrm{E}[X] = \sum_{j\in\mathbb{N}} x_j \Pr(X = x_j). \tag{5.61}$$

The intuition is that

$$\mathrm{E}[X] = \frac{1}{N(\Omega)}\sum_{j\in\mathbb{N}} x_j N(x_j) \tag{5.62}$$

which is precisely what we have.

5.22. Example. Recall Example 3.15 when a student guesses on a true-false exam. What's the expected value of a student guessing on a 10 question true-false exam?

Solution: We see that

$$\mathrm{E}[X] = \sum_{n=0}^{10} n\cdot\binom{10}{n}2^{-10} \tag{5.63}$$

We recall

$$(1+x)^n = \sum_{k=0}^{n}\binom{n}{k}x^k \tag{5.64}$$

thus taking its derivative gives us

$$n(1+x)^{n-1} = \sum_{k=0}^{n} k\binom{n}{k}x^{k-1}. \tag{5.65}$$

We set $x = 1$ and obtain

$$n2^{n-1} = \sum_{k=0}^{n} k\binom{n}{k}. \tag{5.66}$$

We thus deduce

$$\mathrm{E}[X] = 2^{-10}\cdot 10\cdot 2^9 = 5. \tag{5.67}$$

An anticlimactic solution: guessing should give a score of 50%.

5.23. Theorem. *The expectation operator satisfies the following properties:*

1. *If $X \geq 0$, then $\mathrm{E}[X]\geq 0$*
2. *For any $a, b\in\mathbb{R}$ we have $\mathrm{E}[aX+bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$*
3. *The random variable $\mathbf{1}$ which takes the constant value 1 satisfies $\mathrm{E}[\mathbf{1}] = 1$.*

Proof. (1) We see that if X takes values $x_j\geq 0$, then $\Pr(X = x_j)\geq 0$, and the product of two positive real numbers is positive. The sum of positive real numbers is itself a positive real number.

(2) Linearity follows immediately:

$$\begin{aligned}
\mathrm{E}[aX+bY] &= \sum_{\omega} a\omega\Pr(X=\omega) + b\omega\Pr(Y=\omega) && (5.68\text{a})\\
&= \sum_{x} ax\Pr(X=x) + \sum_{y} by\Pr(Y=y) && (5.68\text{b})\\
&= a\sum_{x} x\Pr(X=x) + b\sum_{y} y\Pr(Y=y) = a\mathrm{E}[X] + b\mathrm{E}[Y]. && (5.68\text{c})
\end{aligned}$$

(3) Obvious, since it becomes a sum of probabilities which must be unity. □

5.4 Joint Distributions

5.24. Definition. Let X, Y be discrete random variables. Their "**Joint Distribution**" $F\colon\mathbb{R}^2\to[0,1]$ is given by

$$F(x,y)=\Pr(X\leq x, Y\leq y) \tag{5.69}$$

and their "**Joint Mass Function**" is

$$f(x,y)=\Pr(X=x, Y=y). \tag{5.70}$$

We sometimes use the notation $f_{X,Y}(x,y)$ to make it clear what the random variables are.

5.25. Definition. We see that X and Y are "**Independent**" if and only if

$$\begin{align} f_{X,Y}(x,y) &= \Pr(X=x, Y=y) \tag{5.71a}\\ &= \Pr(X=x)\Pr(Y=y) \tag{5.71b}\\ &= f_X(x)f_Y(y) \tag{5.71c}\end{align}$$

Note we induce this notion using the usual notion of independence.

5.26. Definition. We have the "**Covariance**" of X and Y be

$$\mathrm{Cov}(X,Y)=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y] \tag{5.72}$$

and the "**Correlation Coefficient**"

$$\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \tag{5.73}$$

Recall we defined the variance $\mathrm{Var}(X)$ in (§5.10).

5.27. Lemma (Independence Condition). *Let X, Y be random variables. They are independent if and only if*

$$\mathrm{Cov}(X,Y)=0. \tag{5.74}$$

Proof. We see that

$$\begin{split}\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y] &= \sum_{x,y} xy\Pr(X=x,Y=y)-x\Pr(X=x)y\Pr(Y=y)\\ &= \sum_{x,y} xy\big(\Pr(X=x,Y=y)-\Pr(X=x)\Pr(Y=y)\big)\end{split} \tag{5.75}$$

but we see independence for joint distributions precisely occurs when

$$\big(\Pr(X=x,Y=y)-\Pr(X=x)\Pr(Y=y)\big)=0 \tag{5.76}$$

for any x and y. □

5.28. Theorem (Cauchy-Schwarz Inequality). *Let X, Y be random variables. Then*

$$\big(\mathrm{E}[XY]\big)^2\leq\mathrm{E}[X^2]\mathrm{E}[Y^2] \tag{5.77}$$

with equality if and only if $\Pr(aX=bY)=1$ *for some* $a,b\in\mathbb{R}$ *(at least one of which is nonzero).*

Proof. We introduce a new random variable

$$Z = aX + bY \tag{5.78}$$

where $a, b \in \mathbb{R}$. Suppose $a \geq 0$. Then

$$0 \leq \mathrm{E}[Z^2] = a^2\mathrm{E}[X^2] + b^2\mathrm{E}[Y^2] - 2ab\mathrm{E}XY. \tag{5.79}$$

We consider the right hand side as a quadratic function in a. When does it have a real root? When

$$B^2 - 4AC \leq 0 \tag{5.80a}$$

or for us

$$4b^2\mathrm{E}[XY]^2 - 4\mathrm{E}[X^2] \cdot b^2\mathrm{E}[Y^2] \leq 0 \tag{5.80b}$$

For nonzero b, we have the desired result immediately.

Observe one real root of Eq (5.80) implies

$$\mathrm{E}[Z^2] = 0 \tag{5.81a}$$

which implies

$$Z = 0 \quad \text{with probability 1.} \tag{5.81b}$$

Thus

$$aX - bY = 0 \tag{5.81c}$$

with probability 1. □

5.5 Conditional Distributions and Expectations

Let X and Y be two discrete random variables on $(\Omega, \mathcal{F}, \Pr)$.

5.29. Definition. The "**Conditional Distribution Function**" of Y given $X = x$, written $F_{Y|X}(-|x)$, is defined by

$$F_{Y|X}(y|x) = \Pr(y \leq Y|X = x) \tag{5.82}$$

for any x such that $\Pr(X = x) > 0$.

The "**Conditional Probability Function**" (or "*Conditional Mass Function*") of Y given $X = x$, written $f_{Y|X}(-|x)$, is defined as

$$f_{Y|X}(y|x) = \Pr(Y = y|X = x) \tag{5.83}$$

5.29.1. Remark. Note this definition implies

$$f_{Y|X} = \frac{f_{Y,X}}{f_X} \tag{5.84}$$

and that X and Y are independent if and only if

$$f_{Y|X} = f_Y. \tag{5.85}$$

This justifies the choice of notation.

5.29.2. Remark. Given $X = x$, we may "define"

$$\begin{aligned}\mathrm{E}[Y|X = x] &\overset{\text{def}}{=} \sum_y y f_{Y|X}(y|x) \\ &= \text{conditional expectation of } Y \text{ given } (X = x)\end{aligned} \tag{5.86}$$

We may think of this as a function of x:

$$\psi(x) = \mathrm{E}[Y|X = x]. \tag{5.87}$$

Thus the expected value may be thought of as a function of X:

$$\psi(X) = \mathrm{E}[Y|X]. \tag{5.88}$$

This is a mildly sloppy abuse of notation.

5.30. Definition. Let $\psi(x) = \mathrm{E}[Y|X = x]$. Then the "**Conditional Expectation**" of Y given X written $\mathrm{E}[Y|X]$ is precisely $\psi(X)$. Note that $\mathrm{E}[Y|X]$ is a random variable.

5.31. Theorem. *Given random variables X, Y as specified, then*

$$\mathrm{E}\big[\mathrm{E}[Y|X]\big] = \mathrm{E}[Y]. \tag{5.89}$$

Proof. This is a classic "follow-your-nose and unravel the definitions" type proof.

$$\begin{aligned}\mathrm{E}\big[\mathrm{E}[Y|X]\big] &= \sum_x f_X(x)\left(\sum_y y f_{Y|X}(y|x)\right) && (5.90\text{a})\\ &= \sum_y y \sum_x f_{X,Y}(x,y) && (5.90\text{b})\\ &= \sum_y y f_Y(y) = \mathrm{E}[Y]. && (5.90\text{c})\end{aligned}$$

We just began with Eq (5.86) and "followed our nose"! □

Note more generally, we have

$$\mathrm{E}\big[\mathrm{E}[Y|X]g(X)\big] = \mathrm{E}[Yg(X)]. \tag{5.91}$$

The proof is simple:

$$\begin{aligned}\mathrm{E}\big[\psi(x)g(x)\big] &= \sum_x \psi(x)g(x)\Pr(X = x) && (5.92\text{a})\\ &= \sum_y y \sum_x \Pr(Y = y|X = x)g(x)\Pr(X = x) && (5.92\text{b})\\ &= \sum_{x,y} yg(x)\Pr(X = x, Y = y) = \mathrm{E}[Yg(X)]. && (5.92\text{c})\end{aligned}$$

5.32. Theorem. *Let $a, b \in \mathbb{R}$, and X, Y, Z be random variables.*

1. $\mathrm{E}[aY + bZ|X] = a\mathrm{E}[Y|X] + b\mathrm{E}[Z|X]$
2. $\mathrm{E}[Y|X] \geq 0$ *if* $Y \geq 0$
3. $\mathrm{E}[1|X] = 1$
4. *If X and Y are independent, then* $\mathrm{E}[Y|X] = \mathrm{E}[Y]$
5. $\mathrm{E}[Yg(X)|X] = g(X)\mathrm{E}[Y|X]$ *where* $g\colon \mathbb{R} \to \mathbb{R}$
6. $\mathrm{E}\big[\mathrm{E}[Y|X,Z]|X\big] = \mathrm{E}[Y|X]$ *and* $\mathrm{E}\big[\mathrm{E}[Y|X]|X,Z\big] = \mathrm{E}[Y|X]$.

6 Binomial Confidence Interval

6.1. Motivation. Suppose we flip a coin n times, and we record n_1 heads. How biased is the coin?

6.2. We could suppose these coin flips are random variables, then write

$$S_n = \frac{X_1 + \cdots + X_n}{n}. \tag{6.1}$$

We let

$$\widehat{p} = \frac{n_1}{n}. \tag{6.2}$$

Then the central limit theorem tells us the bias lies in the interval

$$\left[\widehat{p} - z\sqrt{\frac{1}{n}\widehat{p}(1-\widehat{p})}, \widehat{p} + z\sqrt{\frac{1}{n}\widehat{p}(1-\widehat{p})}\right] \tag{6.3}$$

A Set Theory

1.1. Overview. This section is just to establish the notation used for set theory. We use a "naive set theory", which — for the author — is really just ZF+GC.

1.2. Global Choice. Recall we have quantifiers $\forall$ and $\exists$. We can also have a quantifier ε which has the form

$$\varepsilon x : P(x) \tag{A.1}$$

and it returns the object x which satisfies the predicate $P(x)$, if one exists. If there is no such x (e.g., $P(x)$ is a contradiction), then it returns an arbitrary object.

This ε operator is called the "**Global Choice Operator**".

1.3. Definition. A "**Set**" is a well-defined collection of "stuff" without duplicates.

1.4. Definition. If X is a set, and x is an object, if x lives in the collection X we write $x \in X$ and call it a "**Element**" or "**Member**" of X. We will write sets using capital Latin letters unless otherwise indicates. If y does not belong to X, we write $y \notin X$.

Note *any type of object* can belong to a set. For example, we can have a set of selected sets, the set of integers (usually denoted $\mathbb{Z}$), etc.

CAUTION: It is illegal ("meaningless") to write $X \in X$ or $X \notin X$.

1.5. Example. The empty collection is a set, denoted $\emptyset$ and called the "**Empty Set**". It is defined by the condition, for any x, we have $x \notin \emptyset$.

1.6. Example. The collection of natural numbers $\mathbb{N} = \{ 1, 2, 3, \dots \}$, the natural numbers with zero $\mathbb{N}_0 = \{ 0, 1, 2, 3, \dots \}$. The integers $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$. These are all sets.

1.7. Non-Example (Universe). Consider $\mathcal{U}$ the well-defined collection of all sets. It is not a set, since it is illegal to write $\mathcal{U} \in \mathcal{U}$. No, the universe is usually something "bigger" than a set (it's a *class*). We usually don't work with classes, and so we won't worry about them. But we'd like to note the collection of all sets $\mathcal{U}$ is called the "**Universe**".

1.8. Definition. Let X and Y be sets. If every element $x \in X$ belongs to Y, and every $y \in Y$ belongs to X, then we say the two sets are "**Equal**" and write $X = Y$.

1.9. Definition. Let X and Y be sets. If every element $x \in X$ belongs to Y, then we write $X \subseteq Y$ and call X a "**Subset**" of Y.

1.10. Theorem. We have $X = Y$ if and only if $X \subseteq Y$ and $Y \subseteq X$.

1.11. Definition. Let Y be a set. A "**Proper Subset**" X of Y consists of a subset that is not equal to Y. That is: $X \subseteq Y$ and $X \neq Y$. We indicate proper subsets by writing $X \subset Y$.

1.12. Examples. Observe we have $\mathbb{N} \subset \mathbb{N}_0 \subset \mathbb{Z}$.

1.13. Example. For any set X, we have $X \subseteq X$ but $X \not\subset X$.

1.14. Definition. Let X be any set. Then the "**Power Set**" of X is the collection $\mathcal{P}(X)$ of subsets $Y \subseteq X$. Note this implies $X \in \mathcal{P}(X)$.

1.15. Proposition. Let X be any set. Then $\emptyset \in \mathcal{P}(X)$ and $X \in \mathcal{P}(X)$.

1.16. Definition. An "**Ordered Tuple**" (a, b) is a pair of mathematical objects a, b. The first slot a is the first component (sometimes called the *first coordinate*).

We write $(a, b) = (x, y)$ if and only if $a = x$ and $b = y$.

More generally, if we have n objects, we can form the ordered n-tuple $(x_1, \dots, x_n)$. Again, equality is defined component-wise.

1.17. Definition. Let X and Y be sets. Then their "**Cartesian Product**" is a set $X \times Y$ consisting of ordered pairs $(x, y) \in X \times Y$ where $x \in X$, $y \in Y$.

Again, we can generalize this to the Cartesian product of any number of sets $X_1 \times \cdots \times X_n$ consisting of ordered n-tuples.

1.18. Definition. A "**Function**" $f\colon X \to Y$ associates to each $x \in X$ precisely one $y \in Y$ usually denoted $y = f(x)$.

Sometimes mathematicians assert functions are sets. Well, a function f is a subset $f \subseteq X \times Y$ with the property for each $x \in X$, we have precisely one ordered pair $(x, y) \in f$.

1.19. Example. Let X be any set. The identity function $\mathrm{id}\colon X \to X$ defined by $\mathrm{id}(x) = x$ is a function on X.

B References

[1] Geoffrey Grimmet and David Stirzaker,
Probability and Random Processes.
Third ed., Oxford University Press, 2006.

[2] Janko Gravner,
Lecture notes on Probability.
UC Davis Math 135A, 2011. Eprint: `math.ucdavis.edu/~gravner`

[3] L.E. Miller,
"Evaluation of Network Reliability Calculation Method."
Eprint `http://www.antd.nist.gov/wctg/netanal/EvalNetRel.pdf`, 2004.

[4] Kyle Siegrist,
"Virtual Laboratories in Probability and Statistics."
Eprint `math.utah.edu/stat/`, accessed November 9, 2012.

[5] Tom E (`mathoverflow.net/users/3200`),
"What's the use of a complete measure?"
Eprint `mathoverflow.net`, January 12, 2010.