In a Nutshell...
Introduction to Probability and Statistics: A Frequentist Perspective

AT Patera, M Yano[1]

September 12, 2014

1 Preamble

In engineering analysis we must often predict the behavior of a system which is very complicated in the underlying details but which nevertheless admits a rather simple description as regards the *distribution* of the outcomes. The flip of a coin is an immensely complicated phenomenon which depends very sensitively on dozens or even hundreds of variables: we can not predict with any accuracy whether any particular flip will result in a head or a tail; however, we can rather easily describe the frequency of heads and tails in some long sequence of experiments. The life of a person is many orders of magnitude more complicated than the flip of a coin: we can certainly not predict (at birth) the lifespan of any given person; however, again, we can easily describe the distribution of lifespans of many people. In this nutshell, we answer the following question: How can we characterize — and make inferences related to — phenomena or systems which are unpredictable in any given "instance," or individual, but quite readily described for an entire population?

We consider in this nutshell the foundations of probability theory and statistical estimation.

> We first introduce the concepts of population, experiment, outcomes, sample space, sampling procedure, sample, and realization. We then define events as sets of outcomes — hence subsets of the sample space: we describe the set operations with which we can manipulate events; we introduce the concepts of mutually exclusive and collectively exhaustive pairs (or sets) of events; and we provide the classical sample-space Venn diagram visualization of outcomes and events.
>
> We next develop a formulation for frequency as a characterization of a (sample of) data: we introduce the number function and subsequently the frequency of an event; we provide an experiment-space Venn diagram visualization of frequency; we develop the rules by which frequencies associated with different events can be related and combined; we define and relate joint (outcome), marginal, and conditional frequencies;

[1]We thank Ms Debra Blanchard for the preparation of the figures.

and we derive Bayes' Theorem. We also discuss the concept of independence and approximate independence.

Finally, we define probabilities (respectively, the rules of probability) as frequencies (respectively, the rules of frequency) in the limit of infinitely many random experiments. We thereby also naturally introduce the concept of statistical estimation: the process by which we estimate probabilities, or parameters related to probabilities, based on a finite number of experiments. We illustrate these concepts through a pedagogical experiment and associated inference questions.

In this nutshell we emphasize the derivation and interpretation of the rules of probability and the connection between probability and statistical estimation. Subsequent nutshells will consider applications of these rules to describe random phenomena, to calculate probabilities of (complex) events of interest, to assess the error in statistical estimates of probabilities and related distribution parameters, and to develop effective (probabilistic) computational approaches to deterministic problems.

Prerequisites: pre-calculus: elementary set theory, univariate functions, and (conceptual) limits; the letter φ.

2 Motivation: Examples

2.1 A Pedagogical Experiment

We design the survey in the box below to be administered to students enrolled in the MIT Mechanical Engineering subject 2.086, "Numerical Computation for Mechanical Engineers." In the survey, we ask the student to circle one and only one answer to each of two questions. The first question, Question 1, is related to background: has the student not taken, or taken, the MIT Mechanical Engineering subject 2.005, "Thermal-Fluids Engineering I"? The second question, Question 2, is related to knowledge: can the student correctly predict the heat transfer rate through a wall, and in particular the dependence of this heat transfer rate on wall thickness? Note Question 2 should be difficult to answer correctly for a student who has seen the material of 2.005, but relatively easy to answer correctly for a student who has seen — and mastered — the material of 2.005.

The administration of the survey to a (single) student is denoted an *experiment* (or perhaps an *observation*, or a *trial*). A set of n experiments — administration of the survey to n students — constitutes a *sample*; n is denoted the *sample size*. It is important to clearly distinguish an experiment — administration of the survey to a single student — from the sample — administration of the survey to n students.

SURVEY
(for background only)

Please answer the two multiple-choice questions below. You need not put your name on the paper: in this exercise, you are an anonymous member of a sample realization.

1. Heat Transfer Background

   Indicate whether you have (*i*) not yet taken the subject 2.005, or (*ii*) already taken the subject 2.005, by circling the appropriate option below. Note that "already taken" should be interpreted as "completed," so if you are *currently* enrolled in 2.005 then you should circle "(*i*) I have *N*ot yet taken 2.005."

   (*i*) I have *N*ot yet taken 2.005.

   (*ii*) I have already *T*aken 2.005.

2. Heat Transfer Knowledge

   Heat is transferred by conduction through a solid wall to ambient outside air, as shown below. The temperature of the inside wall is fixed at T_{inside}. The thermal conductivity of the wall is denoted k and thickness of the wall is denoted L. The (uniform) heat transfer coefficient from the outside of the wall to the air is given by h and the air temperature far from the wall is T_{amb}. It is known that the Biot number, hL/k, is *very* small (e.g., 10^{-3}). You may treat the problem as one-dimensional: the temperature in the wall varies only in the x-direction.

   If we increase the wall thickness by a factor of two (i.e., from L to $2L$), the heat transfer rate (measured in, say, Watts) through the wall (e.g., from the inside to the outside if T_{inside} is greater than T_{amb}) will

   (*i*) *decrease* by roughly a factor of 2;

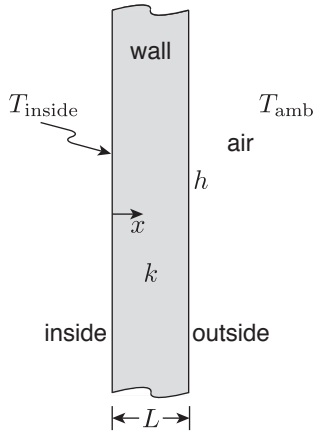
   (*ii*) stay roughly the same;

   (*iii*) *increase* by roughly a factor of 2

   (*iv*) *increase* by roughly a factor of 4.

   Note we only change L to $2L$ — all else (k, h, T_{inside}, T_{amb}) remains fixed.

   Circle one and only one option.



We shall represent our pedagogical experiment by two (dependent) *variables*: one variable related to $\mathbb{B}$(ackground), $\mathbb{B}$, associated to Question 1; the second variable related to $\mathbb{K}$(nowledge), $\mathbb{K}$, associated to Question 2. We may thus express our experiment by a vector, $\mathbb{H}$(eat Transfer Information) $\equiv (\mathbb{B}, \mathbb{K})$: the first element of the pair is the $\mathbb{B}$ackground variable associated with Question 1, and the second element of the pair is the $\mathbb{K}$nowledge variable associated with Question 2. We may think of $\mathbb{B}$ and $\mathbb{K}$ as procedures which yield outcomes B and K, respectively.

For Question 1, represented by variable $\mathbb{B}$, we encode the response (i) as $B = N$, or simply N, for "N(ot taken)"; we encode the response (ii) as $B = T$, or simply T, for "T(aken)." For Question 2, represented by variable $\mathbb{K}$, we encode the response (ii) — the correct answer — as $K = R$, or simply R, for "R(ight)," we encode any other response, (i), (iii), or (iv) — the distractors — as $K = W$, or simply W, for "W(rong)." Since the experiment is described by two variables, and each variable can take on two values, there are four possible *outcomes* to our experiment: $(B, K) = (N, R)$, $(B, K) = (N, W)$, $(B, K) = (T, R)$, and $(B, K) = (T, W)$. The outcome (N, R) is read as "the student has Not taken 2.005" and "the student provides the Right answer to the heat transfer question"; the outcome (N, W) is read as "the student has Not taken 2.005" and "the student provides the Wrong answer to the heat transfer question"; the outcome (T, R) is read as "the student has Taken 2.005" and "the student provides the Right answer to the heat transfer question; and finally, the outcome (T, W) is read as 'the student has Taken 2.005" and "the student provides the Wrong answer to the heat transfer question."

We shall denote our possible outcomes, in general, as $O_1, \ldots, O_{n^{\text{outcomes}}}$. In our experiment — which we shall denote "pedagogical experiment" — $n^{\text{outcomes}} \equiv 4$, and $O_1 \equiv (N, R), O_2 \equiv (N, W), O_3 \equiv (T, R), O_4 \equiv (T, W)$. We denote the set of all possible outcomes — all the possible values for the outcome (B, K) — as the *sample space*: $\{O_1, \ldots, O_{n^{\text{outcomes}}}\}$, or (for our pedagogical example) $\{(N, R), (N, W), (T, R), (T, W)\}$. We emphasize that an experiment will yield *one, and only one*, outcome from the sample space.

It is important to clearly distinguish an experiment from a realization. An *experiment* $\mathbb{H} \equiv (\mathbb{B}, \mathbb{K})$ is the procedure for administration of the survey to any student; an experiment may yield any outcome in our sample space. A *realization* is the administration of the survey to a particular STUDENT*; a realization will yield a single outcome $H^* = (B^*, K^*)$, say $H^* = (N, R)$ — the particular STUDENT* to whom the survey is administered has Not taken 2.005 and (yet!) provides the Right answer to the heat transfer question. In a similar fashion, we can distinguish a sample from a *sample realization*. A sample realization — administration of the survey to n particular students, labeled by number — yields data: n outcomes $\{H_1, \ldots, H_n\}$. We depict in Figure 1 the experiment, a realization and associated outcome, and finally the sample space.

We elaborate briefly on this distinction between an experiment and a realization. Consider a function $f(x) \equiv x^2$: given any real number x, the function f returns an outcome which is a real number. The function f is a *rule* or *procedure* by which, for any given x, we may evaluate $f(x)$. In contrast, for a particular x^*, $y^* \equiv f(x^*)$ is a *particular outcome*

— a *real number*. In the same fashion, $\mathbb{H}$ is a *rule* or *procedure* by which, for any given `student`, we may evaluate $\mathbb{H}(\textsc{Student}) = (\mathbb{B}(\textsc{Student}), \mathbb{K}(\textsc{Student}))$ to yield some outcome from our sample space; in our case, this rule is the administration of our survey. In contrast, a realization is the application of our procedure to a particular $\textsc{Student}^*$ to yield a single outcome $H^* = (B^*, K^*) = (\mathbb{B}(\textsc{Student}^*), \mathbb{K}(\textsc{Student}^*))$ from our sample space. Alternatively, we may develop a programming analogy: the procedure $\mathbb{H}(\textsc{Student})$ is a set of instructions to be parsed and evaluated in terms of a dummy argument Student; a realization, $\mathbb{H}(\textsc{Student}^*)$, is a "call" to the code — which yields *data* H^*.

A student must be enrolled in Mechanical Engineering subject 2.086 to be eligible for the survey; we denote this pool of candidates our *population*. In general, members of a population — here, potential survey respondents — will share some characteristic that defines the population — here, enrollment in 2.086 — but also exhibit some heterogeneity which we wish to investigate — here, as represented by our four possible outcomes. Our population can in principle include students from many different academic years.[2] From this popula-tion we draw our sample of n experiments: n students to which we administer the survey. We might choose to define our sample as all 2.086 students in a given academic year, say $\{H_1, \ldots, H_{n_{2013}}\}_{2013}$; note H_i refers to the outcome for student "i" in the respective sample realizations. Alternatively, we might — and we did — administer the survey to a particular subset of MIT Mechanical Engineering students in a given academic year.

We show in Table 1 an actual sample realization: the results of the survey administered to $n = 64$ distinct students — labeled $1, \ldots, 64$ — who take 2.086 in the Spring of 2013 and furthermore attend lecture on Valentine's Day. (The survey is administered early enough in the semester such that students currently enrolled in 2.005 will have not yet seen the material tested in the heat transfer question and hence rightfully qualify as N rather than T in Question 1.) Note that the order in Table 1, and hence the student labels, are obtained (effectively) by shuffling the responses and then proceeding from the top to the bottom of the resulting pile of sheets.[3] We shall denote the particular sample realization of Table 1 the "Spring2013 Dataset."

We may now ask the following inference questions. Note here "inference" refers to the process by which we deduce conclusions or test hypotheses about our population based on analysis of a sample realization. By way of preamble, we note that if the members of some group of students simply guess the answer to the heat transfer question (Question 2 of the Survey) then we would expect that roughly 25% of the group would obtain the correct answer. We then pose the two inference questions

IQ1 Do the students who have not yet taken 2.005 perform better on the heat transfer

[2] In order to consider this larger population, defined over many years, we must assume that 2.005 is roughly invariant in time. We must also assume that the MIT Mechanical Engineering curriculum — the requirements and the pre- and co-requisite structure — remains unchanged from year to year. But note that neither of these assumptions, in practice not strictly satisfied, affects our analysis of the data for Spring 2013, or our frequentist motivation and derivation of probability theory based on this pedagogical experiment.

[3] In actual fact, the data is compiled in an ordered fashion and we then randomize with the MATLAB built-in `permrand`. It is important to note that we invoke `permrand` *once* and do not in any way attempt to influence the result for (misdirected) purposes of pedagogical point.

$\mathbb{H} = (\mathbb{B}, \mathbb{K})$ EXPERIMENT

SURVEY
Question 1
Question 2
$\left(\text{STUDENT}\right)$

particular $\downarrow$ STUDENT* REALIZATION

$H^* = (B^*, K^*)$ OUTCOME

B: heat transfer Background (Question 1)

K: heat transfer Knowledge (Question 2)

OUTCOMES		B	
		N: heat transfer subject 2.005 Not taken	T: heat transfer subject 2.005 Taken
K	R: Right answer to heat transfer question	(N, R) outcome O_1	(T, R) outcome O_3
	W: Wrong answer to heat transfer question	(N, W) outcome O_2	(T, W) outcome O_4

$\{(N, R), (N, W), (T, R), (T, W)\}$ SAMPLE SPACE

Figure 1: The pedagogical experiment $(\mathbb{B}, \mathbb{K})$, a realization and outcome, and the sample space of possible outcomes.

Student	Outcome	Student	Outcome	Student	Outcome	Student	Outcome
1	(T,W)	2	(N,R)	3	(T,W)	4	(T,W)
5	(N,R)	6	(T,R)	7	(N,W)	8	(T,R)
9	(T,W)	10	(T,R)	11	(N,W)	12	(T,W)
13	(N,W)	14	(N,W)	15	(N,W)	16	(N,W)
17	(T,W)	18	(T,R)	19	(N,R)	20	(N,W)
21	(N,R)	22	(T,W)	23	(T,W)	24	(N,R)
25	(N,W)	26	(T,R)	27	(N,W)	28	(N,W)
29	(N,W)	30	(T,R)	31	(N,W)	32	(T,R)
33	(N,W)	34	(N,W)	35	(T,W)	36	(T,W)
37	(T,W)	38	(N,W)	39	(N,R)	40	(N,W)
41	(T,W)	42	(N,W)	43	(T,W)	44	(N,W)
45	(N,W)	46	(T,W)	47	(N,W)	48	(T,W)
49	(N,W)	50	(T,W)	51	(T,R)	52	(N,R)
53	(T,W)	54	(T,R)	55	(T,R)	56	(T,W)
57	(T,W)	58	(N,R)	59	(T,R)	60	(T,R)
61	(T,W)	62	(N,W)	63	(T,R)	64	(T,R)

Table 1: Sample Realization "Spring2013 Dataset": $\{H_1, \ldots, H_{n=64}\}_{\text{Spring2013}}$.

question than a group of guessers?

IQ2 Do the students who have already taken 2.005 perform better on the heat transfer question than a group of guessers?

We also pose an inference question which directly compares the students who have not yet taken 2.005 with the students who have already taken 2.005:

IQ3 Do students who have taken 2.005 perform better on the heat transfer question than students who have not yet taken 2.005?

Note we can also pose these questions as hypotheses which we wish to accept or reject based on the data obtained.

Our inference questions are not really Yes or No propositions. A student who has dozed through 2.005 and received a well-deserved D(ozed) probably knows less heat transfer than a student who has not taken 2.005 but has worked one summer as an energy auditor. How do we develop a response to our inference question which is useful — provides some insight into the efficacy of 2.005 — but also reflects the *variations* within the Spring 2013 2.086 student body — the literally thousands of factors which contribute to a student's answer to Question 2?

It will require some effort even to answer our inference questions for the particular students associated with the Spring2013 Dataset. But in fact, our goal is more expansive: to

judge the efficacy of 2.005 not just for the students in our particular Spring2013 Dataset, but for the entire larger population of 2.086 students, present, past, and future. This then raises questions about the proper choice of n — the sample size — and indeed the proper choice of students for our realization — the *sampling procedure*. We will discuss sample size, a question of mathematical probability and statistics, in subsequent nutshells, in particular related to the performance of certain estimation and numerical procedures. We shall discuss sampling procedure, a more subtle question of both mathematical and applied statistics, only rather superficially. (We assume for now only that the n experiments in a given realization correspond to n *distinct* students.)

In the exposition that follows, we shall consider this pedagogical experiment — and the particular Spring2013 Dataset sample realization provided in Table 1 — as the primary vehicle by which to develop and illustrate the rules of frequency (and subsequently, probability). (We refer to this experiment as "pedagogical" because of the focus of the experiment and subsequent inferences, not because of the expositional role the example plays in our nutshell.) In most cases we shall also indicate the generalization of our exposition to any {population, experiment, sample space}.

2.2 Some Games of Chance

We introduce as examples several "games of chance" — a flipping coin(s) experiments and a rolling die experiment — on the basis of which you will be able to confirm your understanding of the material as we proceed. These particular experiments are chosen, as in all introductory treatments of probability, because they are simple to understand and "realize" and because the anticipated limiting behavior can be deduced by simple symmetry arguments (though the latter, in fact, belie much underlying complexity, much studied in the academic literature).

We first consider the flipping of a single coin. (Coins shall appear in several, sometimes related, questions throughout this nutshell; we shall assume that in all cases we consider the same denomination — say, a quarter.) More precisely, we define our experiment to be a single flip of a particular coin. Thus in this case we represent the experiment by a single dependent variable, $\mathbb{F}$ (outcome F) which indicates which face of the coin lands "up." We may think of $\mathbb{F}$ as a flipping procedure which takes as argument the time of the flip — a convenient label for an experiment. Our experiment can yield one of two outcomes, $F = T(\text{ail})$ or $F = H(\text{ead})$. It follows that our sample space — the set of all possible outcomes — is given by $\{T, H\}$: $n^{\text{outcomes}} = 2$, and $O_1 \equiv T, O_2 \equiv H$. If we were to flip this same coin n times then we would create a sample of size n: for example, for $n = 3$, we might obtain the sample realization $\{F_1, F_2, F_3\} \equiv \{H, H, T\}$.

CYAWTP 1. Now consider the flipping of two coins. In particular, we define our experiment to be the simultaneous flip of two coins, say one launched from your left thumb, one launched from your right thumb. How many variables do we need to describe the outcome of our experiment? Introduce appropriate name(s) for the variable(s). How many values can each of these variables take on (for a given experiment)? Introduce appropriate name(s) for the

value(s). What are the possible outcomes and hence the sample space for this experiment? Explicitly indicate n^{outcomes} and $\{O_1, O_2, \ldots, O_{n^{\text{outcomes}}}\}$. Now consider a sample of size n (corresponding to n experiments): how many single coin flips — count each thumb as a separate flip — are required to create this sample? Create a sample realization of size $n = 50$.

We next turn to the roll of a die (the singular of dice). We define our experiment to be the roll of a single (standard, six-sided) die. In this case we represent the experiment by a single variable, $\mathbb{D}$ (outcome D), which indicates the number of Dots on that face of the die which lands "up." We may think of $\mathbb{D}$ as a rolling procedure which takes as argument the time of the roll — a convenient lable for an experiment. Our experiment will yield on of six outcomes, $D = 1, D = 2, D = 3, D = 4, D = 5$, and $D = 6$. It follows that our sample space — the set of all possible outcomes — is given by $\{1, 2, 3, 4, 5, 6\}$: $n^{\text{outcomes}} = 6$, and $O_1 \equiv 1, O_2 \equiv 2, O_3 \equiv 3, O_4 \equiv 4, O_5 \equiv 5, O_6 \equiv 6$. If we were to roll the same die n times then we would create a sample of size n: for example, for $n = 5$, we might obtain the sample realization $\{5, 4, 1, 4, 6\}$.

We note that the outcomes of our pedagogical experiment correspond to two variables each of which can take on two values. However, in general, the outcome of an experiment may be described by any number of variables, each of which may take on any number of values (either logical or numerical). In the flip of of a coin the outcomes correspond to one variable which takes on two values; in the roll of a die the outcomes correspond to one variable which takes on six values. The theory we describe can readily accommodate any number of variables, each of which can take on any number of values (or "levels").

3 Events

Given any {population, experiment, sample space}, an *event* $\mathcal{E}$ is, quite simply, a subset of the sample space. We directly illustrate the concept as a sample-space Venn diagram in Figure 2 for the simple case in which $n^{\text{outcome}} \equiv 4$. The event $\mathcal{E}_1$ is the set $\{O_1, O_2\}$; the event $\mathcal{E}_2$ is the set $\{O_1, O_3\}$. Note that the dashed lines which enclose the outcomes associated with $\mathcal{E}_1$ and $\mathcal{E}_2$ are schematic: a graphical version of $\{ \cdot \}$. If the result of an experiment is an outcome which belongs to the set $\mathcal{E}$ (respectively, does not belong to the set $\mathcal{E}$) we say that "event $\mathcal{E}$ happens" or "event $\mathcal{E}$ occurs" (respectively, "event $\mathcal{E}$ does not happen" or "event $\mathcal{E}$ does not occur"). We consider our simple example of Figure 2: if the experiment yields outcome O_1 then event $\mathcal{E}_1$ happens and also event $\mathcal{E}_2$ happens; if the experiment yields outcome O_2 then event $\mathcal{E}_1$ happens but event $\mathcal{E}_2$ does not happen; if the experiment yields outcome O_3 then event $\mathcal{E}_1$ does not happen but event $\mathcal{E}_2$ does happen; and finally, if the experiment yields O_4 then event $\mathcal{E}_1$ does not happen and also event $\mathcal{E}_2$ does not happen. Alternatively, from an event perspective, if an experiment yields either outcome O_1 or outcome O_2, then $\mathcal{E}_1$ happens; conversely, if an experiment yields neither O_1 nor O_2, then $\mathcal{E}_1$ does not happen. Similarly for $\mathcal{E}_2$: if an experiment yields either outcome O_1 or outcome O_3, then $\mathcal{E}_2$ happens; conversely, if an experiment yields neither O_1 nor O_4, then $\mathcal{E}_2$ does not happen.

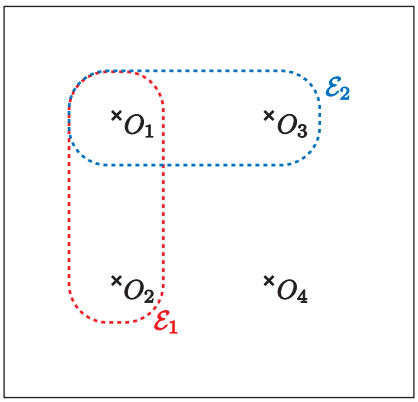


Figure 2: A Venn diagram representation of the sample space of ($n_{\text{outcomes}} = 4$) possible outcomes, $\{O_1, O_2, O_3, O_4\}$, and two events $\mathcal{E}_1$ (red dashed line) and $\mathcal{E}_2$ (blue dashed line).

We now recall the elementary set operations which shall prove useful in the definition and manipulation of events.

Complement. We denote by $\bar{\mathcal{E}}$ the *complement* of $\mathcal{E}$ in the sample space: $\bar{\mathcal{E}}$ is all the outcomes in the sample space which do *not* belong to $\mathcal{E}$. We note that if event $\mathcal{E}$ happens then event $\bar{\mathcal{E}}$ does not happen — since by construction $\mathcal{E}$ and $\bar{\mathcal{E}}$ do not share any outcomes.

Union. We denote by $\mathcal{E}_1 \cup \mathcal{E}_2$ the *union* of two events $\mathcal{E}_1$ and $\mathcal{E}_2$: event $\mathcal{E}_1 \cup \mathcal{E}_2$ is the set of all outcomes which appear in either $\mathcal{E}_1$ *or* $\mathcal{E}_2$; event $\mathcal{E}_1 \cup \mathcal{E}_2$ happens if *either* $\mathcal{E}_1$ happens *or* $\mathcal{E}_2$ happens.

Intersection. We denote by $\mathcal{E}_1 \cap \mathcal{E}_2$ the *intersection* of two events $\mathcal{E}_1$ and $\mathcal{E}_2$: event $\mathcal{E}_1 \cap \mathcal{E}_2$ is the set of all outcomes which appear in both $\mathcal{E}_1$ *and* $\mathcal{E}_2$; event $\mathcal{E}_1 \cap \mathcal{E}_2$ happens if *both* $\mathcal{E}_1$ happens *and* $\mathcal{E}_2$ happens.

We provide an illustration of each of the above for our example of Figure 2. The complement of $\mathcal{E}_1$, $\bar{\mathcal{E}}_1$, is given by $\bar{\mathcal{E}}_1 \equiv \{O_3, O_4\}$. The union of $\mathcal{E}_1$ and $\mathcal{E}_2$, $\mathcal{E}_3 \equiv \mathcal{E}_1 \cup \mathcal{E}_2$, is given by $\mathcal{E}_3 \equiv \{O_1, O_2, O_3\}$. The intersection of $\mathcal{E}_1$ and $\mathcal{E}_2$, $\mathcal{E}_4 \equiv \mathcal{E}_1 \cap \mathcal{E}_2$, is given by $\mathcal{E}_4 \equiv \{O_1\}$. The set operations complement, union, and intersection correspond to the logical operations "not," "or," and "and," respectively; the latter are convenient in reading and interpreting events.

CYAWTP 2. Consider the simple example of Figure 2. What outcomes belong to the event (set) $\overline{\mathcal{E}_1 \cup \mathcal{E}_2}$?

Let us know consider any two events, $\mathcal{E}_1$ and $\mathcal{E}_2$ (not necessarily the events of Figure 2). If $\mathcal{E}_1 \cap \mathcal{E}_2 = \emptyset$ (the empty set), we say that events $\mathcal{E}_1$ and $\mathcal{E}_2$ are *mutually exclusive*. In words, $\mathcal{E}_1$ and $\mathcal{E}_2$ are mutually exclusive if they share no outcomes; equivalently, $\mathcal{E}_1$ and $\mathcal{E}_2$ are mutually exclusive if "$\mathcal{E}_1$ happens" *implies* "$\mathcal{E}_2$ does not happen" (and of course also the

converse) — $\mathcal{E}_1$ and $\mathcal{E}_2$ can not *both* happen. As already indicated, for any event $\mathcal{E}$, $\mathcal{E}$ and $\bar{\mathcal{E}}$ are mutually exclusive. A *set* of M events $\mathcal{E}_m, 1 \leq m \leq M$, is mutually exclusive if each pair of events in the set of events is mutually exclusive: no two events in the set of events share an outcome; equivalently, in any given experiment, at most only one event in the set of events can happen. In our particular example of Figure 2, the events $\mathcal{E}_1$, $\{O_3\}$, and $\{O_4\}$ are mutually exclusive, however the events $\mathcal{E}_1$ and $\mathcal{E}_2$ are not mutually exclusive.

Let us again consider any two events, $\mathcal{E}_1$ and $\mathcal{E}_2$ (not necessarily the events of Figure 2). If $\mathcal{E}_1 \cup \mathcal{E}_2$ is the entire sample space, we say that $\mathcal{E}_1$ and $\mathcal{E}_2$ are *collectively exhaustive*. In words, $\mathcal{E}_1$ and $\mathcal{E}_2$ are collectively exhaustive if all possible outcomes of the experiment appear in either $\mathcal{E}_1$ or $\mathcal{E}_2$; equivalently, in any given experiment, either $\mathcal{E}_1$ or $\mathcal{E}_2$ *must* happen. For any event $\mathcal{E}$, $\mathcal{E}$ and $\bar{\mathcal{E}}$ are collectively exhaustive. A *set* of M events $\mathcal{E}_m, 1 \leq m \leq M$, is collectively exhaustive if the union of all the events in the set is the entire sample space; equivalently, in any given experiment, at least one of the events in the set of events must happen. In our particular example of Figure 2, the events $\mathcal{E}_3 \equiv \mathcal{E}_1 \cup \mathcal{E}_2$, $\{O_3\}$, and $\{O_4\}$ are collectively exhaustive, however the events $\mathcal{E}_1$ and $\mathcal{E}_2$ are not collectively exhaustive.

A pair of events, or a set of events, may be both mutually exclusive and collectively exhaustive. In this case, the result of an experiment is one event, and only one event, in the set of events: we know that at most one event can happen since the set of events is mutually exclusive; we note that at least one event must happen since the set of events is collectively exhaustive. An important case is the set of n^{outcome} events $\{O_1\}$, $\{O_2\}$, ..., $\{O_{n^{\text{outcomes}}}\}$. Clearly this set of events is both mutually exclusive and collectively exhaustive: an experiment can yield only one outcome, hence mutually exclusive; an experiment must yield at least one outcome, hence collectively exhaustive. The pair of events $\mathcal{E}$ and $\bar{\mathcal{E}}$ (for any event $\mathcal{E}$) is another important example of a set of mutually exclusive and collectively exhaustive events. Finally, the event $\emptyset$ — the empty set — and the sample space are yet another example of mutually exclusive and collectively exhaustive events.

To close this section, we consider events associated with our pedagogical experiment. We first introduce a shorthand. We shall denote the event $\mathcal{E} \equiv \{(N, R), (N, W)\}$ by N: the event N is the set of outcomes $H = (B, K)$ for which $B = N$ (and K may take on any value). In words, the event N reads "the student has *N*ot taken 2.005": if a particular realization of the experiment (hence student) yields outcome $B = N$ (and any value for K), N happens — the student *has not taken* 2.005; conversely, if the result of the experiment is outcome $B = T$ (and any value for K), the event N does not happen — the student *has not not taken* 2.005, or equivalently the student *has taken* 2.005. In a similar fashion we define events T ("the student has *T*aken 2.005"), R ("the student provides the *R*ight answer to the heat transfer question"), and W ("the student provides the *W*rong answer to the heat transfer question").

The Venn diagram of Figure 2 provides a visualization of the sample space associated with our pedagogical experiment: we identify $O_1 \equiv (N, R), O_2 \equiv (N, W), O_3 \equiv (T, R), O_4 \equiv (T, W)$. We observe that the event $\mathcal{E}_1$ is the event N: "the student has not *T*aken 2.005." Similarly, the event $\mathcal{E}_2$ is the event R: "the student provides the *R*ight answer to the heat transfer question." The event $\mathcal{E}_1 \cap \mathcal{E}_2$ is the event $O_1 \equiv (N, R)$: "the student has *N*ot taken

2.005" and "the student provides the *R*ight answer to the heat transfer question." We can immediately conclude from the Venn diagram that the events N (the leftmost two events) and T (the rightmost two events) are mutually exclusive and collectively exhaustive; the student must either *not have taken* or *have taken* 2.005. Similarly, the events R and W are mutually exclusive and collectively exhaustive.

CYAWTP 3. Reconsider the experiment of simultaneously flipping two coins first introduced in **CYAWTP 1**. Define two events which are mutually exclusive but not collectively exhaustive. Define two events which are collectively exhaustive but not mutually exclusive. Define two events which are both mutually exclusive and collectively exhaustive. Define two events which are neither mutually exclusive nor collectively exhaustive. In all cases express each event in words but also more explicitly as the associated set of outcomes.

4 Frequencies

4.1 The "Number" Function

We now define a "number" function $\#(\mathcal{E})$. We are given some {population, experiment, sample space}, an event $\mathcal{E}$, and a sample realization of size n: $\#(\mathcal{E})$ is then the number of occurrences of event $\mathcal{E}$ in the sample (or sample realization); more explicitly, $\#(\mathcal{E})$ is the number of experiments which yield an outcome which is in the set of outcomes which defines $\mathcal{E}$.

We will illustrate the concept in the context of our pedagogical experiment and the sample realization Spring2013 of Table 1. We first consider the event $\mathcal{E} \equiv \{(N, R)\}$: the student has not taken 2.005 ($B = N$) and has provided the right answer to the heat transfer question ($K = R$). We conclude from Table 1 that, for realization Spring2013, $\#(\mathcal{E}) = 8$: there are 8 students for which this event happens. We can justify this conclusion in several equivalent fashions: there are 8 experiments in Spring2013 Dataset for which B takes on the value N and K takes on the value R — (N, R) appears in 8 entries of Table 1; more formally, there are 8 experiments in Spring 2013 Dataset for which the outcome is in the set $\mathcal{E} \equiv \{(N, R)\}$. (Of course, for our singleton set, an outcome is in $\{(N, R)\}$ if and only if the outcome is exactly (N, R).) We next consider the event $\mathcal{E} \equiv N$, which we recall is the event "the student has not taken 2.005." We determine from Table 1 that, for realization Spring2013, $\#(\mathcal{E}) = 30$: there are 30 students for which the event happens. The justification: there are 30 experiments in Spring2013 Dataset for which B takes on the value N — N appears in 30 entries of Table 1; more formally, there are 30 experiments in Spring2013 Dataset for which the outcome is in the set $\mathcal{E} \equiv N \equiv \{(N, R), (N, W)\}$ — the experiment yields as outcome either (N, R) or (N, W). We tabulate in Table 2 $\#(\mathcal{E})$ for several important events.

Returning to the case of a general {population, experiment, sample space}, it will be convenient to develop certain relationships which will, in turn, permit us to easily calculate the "number" function associated with two events, say $\mathcal{E}_1$ and $\mathcal{E}_2$. To begin, we introduce a picture — another Venn diagram — which will often help us understand the underlying set operations which inform the "number" calculation. In particular, we depict in Figure 3

$\#(N,R)$	$\#(N,W)$	$\#(T,R)$	$\#(T,W)$	$\#(N)$	$\#(T)$	$\#(R)$	$\#(W)$
8	22	14	20	30	34	22	42

Table 2: The number of occurrences in the Spring2013 Dataset of some important events associated with our pedagogical experiment.

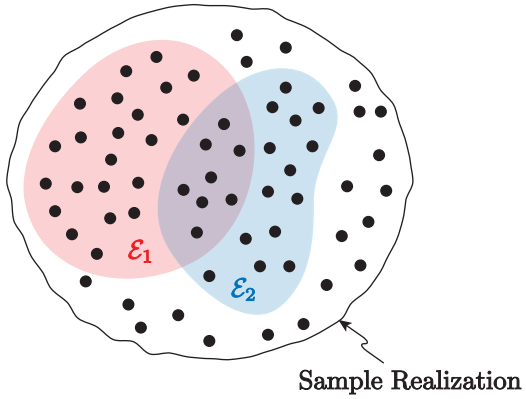


Figure 3: A Venn diagram representation of the number function evaluated for events $\mathcal{E}_1$ (red blob) and $\mathcal{E}_2$ (blue blob) over a sample realization of $n = 64$ experiments (black dots).

our sample realization as a disk within which we place a dot representing the outcome of each of the n individual experiments. (In our pedagogical experiment, each dot represents a student.) We denote by the red blob the subset of the sample realization for which event $\mathcal{E}_1$ happens: the red blob contains the black dots associated with those experiments for which the outcome is in the set $\mathcal{E}_1$. We denote by the blue blob the subset of the sample realization for which event $\mathcal{E}_2$ happens: the blue blob contains the black dots associated with those experiments for which the outcome is in the set $\mathcal{E}_2$. The bounding black line, which includes all the black dots, represents our sample realization. (Of course the depiction is schematic, and thus we are free to assume that the blobs form smooth connected domains. Any shapes which contain the requisite dots will suffice.)

We emphasize that the Venn diagram of Figure 2 and the Venn diagram of Figure 3 are fundamentally different: the former is in "outcome space" with each "x" representing an outcome; the latter is in "experiment space," with each dot representing an experiment (in our pedagogical experiment, a student). The dashed lines in Figure 2 enclose a set of outcomes which define an event. The blobs in Figure 3 enclose a set of experiments in particular sample realization for which an event happens.

CYAWTP 4. (*a*) Modify Figure 3 for the case in which $\mathcal{E}_1$ and $\mathcal{E}_2$ are mutually exclusive but not collectively exhaustive. (*b*) Modify Figure 3 for the case in which $\mathcal{E}_1$ and $\mathcal{E}_2$ are not mutually exhaustive but are collectively exhaustive. (*c*) Modify Figure 3 for the case in which $\mathcal{E}_1$ and $\mathcal{E}_2$ are mutually exclusive and collectively exhaustive.

CYAWTP 5. The depiction of Figure 3 is generic but in fact also corresponds, in quantitative fidelity, to our pedagogical experiment: sample realization Spring2013 Dataset; events $\mathcal{E}_1 \equiv N$ and $\mathcal{E}_2 \equiv R$. Provide a sketch analogous to Figure 3, again for sample realization Spring 2013 Datset, for the events $\mathcal{E}_1 \equiv T$ and $\mathcal{E}_2 \equiv R$.

We now demonstrate various relations for the number function by the well-known "method of dots." To begin, we refer to your figure of **CYAWTP 4** (c) to directly conclude that, for two mutually exclusive and collectively exhaustive events,

$$\#(\mathcal{E}_1) + \#(\mathcal{E}_2) = n \text{ (} \textbf{if } \mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ are mutually exclusive and collectively exhaustive)}; \quad (1)$$

as a special case, we conclude that, for any event $\mathcal{E}$,

$$\#(\mathcal{E}) + \#(\bar{\mathcal{E}}) = n \ , \quad (2)$$

since $\mathcal{E}$ and $\bar{\mathcal{E}}$ are perforce mutually exclusive and collectively exhaustive.

Let us say that we now wish to calculate $\#(\mathcal{E}_1 \cup \mathcal{E}_2)$: the number of experiments in our sample for which either $\mathcal{E}_1$ happens or $\mathcal{E}_2$ happens. The appropriate number of dots (in our pedagogical experiment, appropriate number of students) is clearly, from inspection,

A [the number of dots in the red blob] **plus** [the number of dots in the blue blob] **minus** [the number of dots in both the red blob *and* the blue blob (in fact, a purple blob)]; or

B [n (the number of black dots in the sample realization)] **minus** [the number of dots not in the red blob *and* not in the blue blob].

We now develop formulas based on the two constructions A and B.

We first consider construction A. To begin, we note that the number of dots in both the red blob and the blue blob (purple blob) — the *intersection* of the red blob and the red blob — is the number of experiments for which both $\mathcal{E}_1$ *and* $\mathcal{E}_2$ happen, or $\#(\mathcal{E}_1 \cap \mathcal{E}_2)$; recall that $\mathcal{E}_1 \cap \mathcal{E}_2$ is the set of outcomes which belong to both $\mathcal{E}_1$ and $\mathcal{E}_2$. Hence we may express construction A as

$$\#(\mathcal{E}_1 \cup \mathcal{E}_2) = \ \#(\mathcal{E}_1) \ + \ \#(\mathcal{E}_2) \ - \ \#(\mathcal{E}_1 \cap \mathcal{E}_2) \ . \quad (3)$$

This form is convenient because it is often relatively easy to evaluate the $\cap$, either because $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent (a concept introduced later in this nutshell) or because $\mathcal{E}_1$ and $\mathcal{E}_2$ are mutually exclusive — the case depicted in your figure of **CYAWTP 4** (a) (or (c)). In the latter case,

$$\#(\mathcal{E}_1 \cup \mathcal{E}_2) = \#(\mathcal{E}_1) \ + \ \#(\mathcal{E}_2) \quad (\textbf{if } \mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ are mutually exclusive}) \ , \quad (4)$$

since $\mathcal{E}_1 \cap \mathcal{E}_2$ is perforce empty for all experiments: no outcome belongs to both $\mathcal{E}_1$ and $\mathcal{E}_2$— the intersection of the red blob and blue blob is perforce empty.[4] The expression (4), but less so (3), extends readily from two events to any number of events.

[4] We note a subtlety: for a particular n and a particular sample realization it is certainly possible that (4) is correct even for two events $\mathcal{E}_1$ and $\mathcal{E}_2$ which are not mutually exclusive: the particular sample realization may not contain any experiments for which the outcome belongs to both $\mathcal{E}_1$ and $\mathcal{E}_2$. However, we wish to derive relationships which are valid for any sample realization.

Next, we consider construction B. We may express this formulation as

$$\#(\mathcal{E}_1 \cup \mathcal{E}_2) = n - \#(\bar{\mathcal{E}}_1 \cap \bar{\mathcal{E}}_2) \ , \tag{5}$$

which in fact derives from De Morgan's laws and (2) — but is readily evident in Figure 3. The form (5) again involves an *cap* and also extends readily from two events to any number of events.

4.2 The "Frequency" Function

We many now introduce a "frequency" function $\varphi_n(\mathcal{E})$. We are given some {population, experiment, sample space}, an event $\mathcal{E}$, and a sample realization of size n. Then

$$\varphi_n(\mathcal{E}) \equiv \frac{\#(\mathcal{E})}{n} \ ,$$

which measures the *fraction* of experiments in our sample realization for which event $\mathcal{E}$ happens. Note that since $0 \le \#(\mathcal{E}) \le n$ it follows that $0 \le \varphi_n(\mathcal{E}) \le 1$.

We illustrate the concept in the context of our pedagogical experiment and the sample realization Spring2013 of Table 1. We consider the event $\mathcal{E} \equiv (N, R)$: we recall from Section 4.1 and Table 2 that $\#(\mathcal{E}) = 8$; it immediately follows that $\varphi_n(\mathcal{E}) = \#(\mathcal{E})/n = 8/64 = 0.125$. (Recall that our sample realization Spring2013 is of size $n = 64$.) We next consider the event $\mathcal{E} \equiv N$: we recall from Section 4.1 and Table 2 that $\#(\mathcal{E}) = 30$; it immediately follows that $\varphi_n(\mathcal{E}) = \#(\mathcal{E})/n = 30/64 \approx 0.469$ — 46.9% of our Spring2013 sample of students has not yet taken 2.005. We shall calculate many other frequencies associated with our pedagogical experiment in the next sections.

We note that the frequency function will inherits the properties described in Section 4.1: we need only divide through each equation by n. We list the results for future reference. First, it follows from (1) that, for two mutually exclusive and collectively exhaustive events $\mathcal{E}_1$ and $\mathcal{E}_2$,

$$\varphi_n(\mathcal{E}_1) + \varphi_n(\mathcal{E}_2) = 1 \ (\textbf{if } \mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ are mutually exclusive and collectively exhaustive}); \tag{6}$$

as expected, if we consider two events which exhaust the sample space, one or the other must happen "all the time" — hence with frequency unity. As a special case, we conclude from (2) that, for any event $\mathcal{E}$,

$$\varphi_n(\mathcal{E}) + \varphi_n(\bar{\mathcal{E}}) = 1 \ , \tag{7}$$

since $\mathcal{E}$ and $\bar{\mathcal{E}}$ are perforce mutually exclusive and collectively exhaustive. Second, it follows from (3) that, for any two events $\mathcal{E}_1$ and $\mathcal{E}_2$,

$$\varphi_n(\mathcal{E}_1 \cup \mathcal{E}_2) = \ \varphi_n(\mathcal{E}_1) \ + \ \varphi_n(\mathcal{E}_2) \ - \ \varphi_n(\mathcal{E}_1 \cap \mathcal{E}_2) \ ; \tag{8}$$

for the case in which $\mathcal{E}_1$ and $\mathcal{E}_2$ are mutually exclusive, we obtain from (4) that

$$\varphi_n(\mathcal{E}_1 \cap \mathcal{E}_2) = \varphi_n(\mathcal{E}_1) \ + \ \varphi_n(\mathcal{E}_2) \quad (\textbf{if } \mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ are mutually exclusive}) \ . \tag{9}$$

Third, it follows from (5) that, for any two events $\mathcal{E}_1$ and $\mathcal{E}_2$,

$$\varphi_n(\mathcal{E}_1 \cup \mathcal{E}_2) = 1 - \varphi_n(\bar{\mathcal{E}}_1 \cap \bar{\mathcal{E}}_2) \; ; \tag{10}$$

recall that to derive (10) we divide both sides of (5) by n, invoke the definition of $\varphi_n(\cdot) = \#(\cdot)/n$, and observe that $n/n = 1$.

Lastly, we consider the case of M events $\mathcal{E}_m, 1 \leq m \leq M$. We first consider the extension of (6): for M events $\mathcal{E}_m, 1 \leq m \leq M$, mutually exclusive and collectively exhaustive,

$$\sum_{m=1}^{M} \varphi_n(\mathcal{E}_m) = 1 \quad (\mathcal{E}_m, 1 \leq m \leq M, \text{ mutually exclusive and collectively exhaustive}) \; . \tag{11}$$

The proof is simple: each dot in our sample realization must appear in at most one event blob, since the events are mutually exclusive; each dot in our sample realization must appear in at least one event blob, since the events are collectively exhaustive. We next consider the extension of (10): for any events $\mathcal{E}_m, 1 \leq m \leq M$, not (necessarily) either mutually exclusive or collectively exhaustive,

$$\varphi_n(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \cdots \cup \mathcal{E}_M) = 1 - \varphi_n(\bar{\mathcal{E}}_1 \cap \bar{\mathcal{E}}_2 \cap \cdots \cap \bar{\mathcal{E}}_M). \tag{12}$$

Note that the intersection of M events contains only those outcomes which appear in each event.

CYAWTP 6. Provide a sketch analogous to Figure 3 and associated "proof by dots" for the relationship (12) for the case of $K = 4$ events.

The approach above is top-down: we begin with certain events, and then deduce relationships between these events. We may also pursue an approach which is more bottom-up: we begin with outcomes, and then express an event in terms of *outcome frequencies*.

We know that an event $\mathcal{E}$ is a set a outcomes (which constitute a subset of the sample space). To be more explicit, we denote $\mathcal{E}$ as

$$\mathcal{E} \equiv \{O_{i_1}, O_{i_2}, \ldots, O_{i_J}\} \; ; \tag{13}$$

here J — the number of outcomes in $\mathcal{E}$ — and the indices $i_j, 1 \leq j \leq J$ — which identify the particular outcomes associated with event $\mathcal{E}$ — will of of course depend on the event $\mathcal{E}$ of interest. We now note that our outcomes are, by definition, mutually exclusive, and hence from (9)

$$\varphi_n(\mathcal{E}) = \sum_{j=1}^{J} \varphi_n(O_{i_j}) \; , \tag{14}$$

where

$$\varphi_n(O_k) = \frac{\#(O_k)}{n} \; , \; 1 \leq k \leq n^{\text{outcomes}} \; . \tag{15}$$

Note that $\#(O_k)$ is the frequency of outcome O_k: the number of experiments in the sample realization which yield outcome O_k. The interpretation of (14) is simple: the number of occurrences of $\mathcal{E}$ in our sample realization is the sum of the number of the occurrences of the outcomes which comprise $\mathcal{E}$.

The formula (14) provides a general procedure for the calculation of frequencies of any event, $\mathcal{E}$, for some given sample realization. First, we evaluate the *outcome frequencies* $\varphi_n(O_k), 1 \le k \le n^{\text{outcomes}}$. Second, we identify J and the indices (relevant outcomes) $i_j, 1 \le j \le J$, for any particular event of interest, $\mathcal{E}$. (In fact, this second step is independent of the realization.) Third, we perform the sum (14).

CYAWTP 7. Consider as our experiment the roll of a single die, as introduced in Section 2.2. Create a sample realization of size $n = 36$. Calculate the outcome frequencies associated with our sample space $\{1, 2, 3, 4, 5, 6\}$. Now define the event $\mathcal{E} \equiv$ "D is odd": the die roll yields an odd number of dots. Identify J and the $i_j, 1 \le j \le k$, associated with this event. Evaluate $\varphi_n(\mathcal{E})$ in two ways, (i) directly from the sample, as $\#(\text{"}D \text{ is odd"})/n$, and (ii) from the formula (14).

4.3 Joint, Marginal, and Conditional Frequencies

4.3.1 Joint Frequencies

We develop the concept of *joint frequencies* within the context of our pedagogical experiment. In particular, the joint frequencies are the frequencies associated with all possible values of the two variables, B and K, which describe the outcome of our pedagogical experiment: $\varphi_n(N, R)$, $\varphi_n(N, W)$, $\varphi_n(T, R)$, and $\varphi_n(T, W)$. We recall that for our pedagogical experiment the sample space is given by $\{O_1 \equiv (N, R), O_2 \equiv (N, W), O_3 \equiv (T, R), O_4 \equiv (T, W)\}$. We thus observe that the joint frequencies are precisely the outcome frequencies of Section 4.2 and in particular (15): $\varphi_n(O_1) \equiv \varphi_n(N, R), \ldots, \varphi_n(O_4) \equiv \varphi_n(T, W)$. This is generally true for any experiment in which the outcomes are represented in terms of (any number of) values of (any number of) variables. We can thus understand the importance of the joint frequencies: a set of frequencies in terms of which we can express, through (14), the frequency of any event $\mathcal{E}$.

We can readily deduce the joint frequencies for the sample realization Spring2013 of Table 1 from the summary of Table 2: $\varphi_n(N, R)$ is given by $\#(N, R)/n = 8/64 = 0.125$; similarly, $\varphi_n(N, W) = 22/64 \approx 0.344$, $\varphi_n(T, R) = 14/64 \approx 0.219$, and $\varphi_n(T, W) = 20/64 \approx 0.313$. We summarize these joint frequencies (to three digits) in Table 3. (We could also present the joint frequencies graphically, as a histogram; we shall introduce histograms in a subsequent nutshell, in particular for the case of many outcomes.) We note from our calculations that

$$\varphi_n(N, R) + \varphi_n(N, W) + \varphi_n(T, R) + \varphi_n(T, W) = 1 \; . \tag{16}$$

We can easily anticipate this result: each experiment (each student) yields at most one outcome — our outcomes are mutually exclusive — and at least one outcome — our outcomes

$\varphi_n(N,R)$ $(\equiv \varphi_n(O_1))$	$\varphi_n(N,W)$ $(\equiv \varphi_n(O_2))$	$\varphi_n(T,R)$ $(\equiv \varphi_n(O_3))$	$\varphi_n(T,W)$ $(\equiv \varphi_n(O_4))$
0.125	0.344	0.219	0.313

Table 3: Joint frequencies (also outcome frequencies) associated with Spring2013 Dataset.

$\varphi_n(N)$	$\varphi_n(T)$	$\varphi_n(R)$	$\varphi_n(W)$
0.469	0.531	0.344	0.656

Table 4: Marginal frequencies associated with Spring2013 Dataset.

are collectively exhaustive; hence each of our n students is counted in at most one and at least one, and hence only one, of $\#(N,R)$, $\#(N,W)$, $\#(T,R)$, or $\#(T,W)$; thus $\#(N,R) + \#(N,W) + \#(T,R) + \#(T,W) = n$ — which, upon division by n, yields (16). Our argument is quite general: (16) is valid for any (two-variable, two-value) experiment and any sample realization.

We note that (16) is just a special case of (11). The relation (11) in fact implies, for any {population, experiment, sample space},

$$\sum_{i=1}^{n^{\text{outcomes}}} \varphi_n(O_i) = 1 \; . \tag{17}$$

since our outcomes constitute a set of "elementary" events which is both mutually exclusive and collectively exhaustive. For an experiment for which the outcomes are represented in terms of (any number of) values of (any number of) variables, the $\varphi_n(O_i)$ are just the joint frequencies.

4.3.2 Marginal Frequencies

We now develop the concept of marginal frequencies, again within the context of our pedagogical experiment. We shall consider B in our exposition; K follows a parallel development.

We recall that, in any outcome (B,K), B may take on two values, N (for Not taken 2.005), or T (for Taken 2.005). We know that we can deduce the frequency of these events, $\varphi_n(N)$ and $\varphi_n(T)$, respectively, directly from our sample realization (and hence, for our particular sample realization Spring2013 Dataset, directly from Table 1. But we can also deduce the frequencies of events $B = \cdot$ from the joint frequencies of (B,K): we shall remove, or marginalize, the effect of K; the resulting frequencies are thus denoted *marginal* frequencies. (Of course, in the case of an experiment described by a single variable, we do not need the adjectives joint or marginal to describe the frequency.)

We first note that the event $B = N$ is equivalent to the event $\{(N,R),(N,W)\}$: in words, we include in the event $B = N$ all the outcomes for which $B = N$ without consideration

of the value of the other variable, K. We can then apply the relation (14) for $J \equiv 2$ and $i_1 \equiv 1, i_2 \equiv 2$ — recall our outcomes are given by $\mathcal{O}_1 \equiv (N,R), \mathcal{O}_2 \equiv (N,W), \mathcal{O}_3 \equiv (T,R), \mathcal{O}_4 \equiv (T,W)$ — to obtain

$$\varphi_n(N) = \varphi_n(N,R) + \varphi_n(N,W) \ , \tag{18}$$

We illustrate the result for our particular sample realization, Spring2013 Dataset of Table 1: we recall from Table 3 that the relevant joint frequencies are given by $\varphi_n(N,R) = 8/64 = 0.125, \varphi_n(N,W) = 22/64 \approx 0.344$; it then follows from (18) that $\varphi_n(N) = (8+22)/64 = 30/64 \approx 0.469$ — just as we derived in Section 4.2 by directly counting events in Table 1. We can similarly calculate all of the marginal frequencies, associated with both B and K, as summarized (to three digits) in Table 4.

CYAWTP 8. Derive expressions for $\varphi_n(T), \varphi_n(R)$, and $\varphi_n(W)$ analogous to (18) for $\varphi_n(N)$ and confirm the numerical values provided in Table 4. Demonstrate that $\varphi_n(N) + \varphi_n(T) = 1$ and also $\varphi_n(R) + \varphi_n(W) = 1$ for any sample realization, and then confirm these relations empirically for the particular sample realization Spring2013.

We can readily extend (18) to an experiment for which the outcomes are described by any number of variables each of which may take on any number of values. In particular, to obtain the marginal frequency for any particular value of a given variable, we simply sum the joint frequencies for the particular value of the given variable over all possible values of all the other variables.

4.3.3 Conditional Frequencies

Definition and Interpretation. We shall first derive expressions for particular conditional frequencies in the context of our pedagogical experiment and the particular Spring2013 sample realization of Table 1. We shall subsequently consider extension to any {population, experiment, sample space}.

Say we wish to understand the effect of N on R: if a student has Not yet taken 2.086, how will this affect the ability of the student to provide the Right answer to the heat transfer question? We proceed in two stages.

1. We are interested in students who have Not yet taken 2.086. We denote by Spring2013′ the subset of the sample realization Spring2013 which contains only those experiments for which the outcome is (N,K) for *any* value of K — in short, $B = N$; we further denote by n' the sample size of Spring2013′. We present Spring2013′ in Table 5; we observe from Table 5 that $n' = 30$. Note that n' will depend on the event on which we condition; in our example here, we "condition on" $B = N$.

2. We are interested in those students who have Not yet taken 2.086 who furthermore provide the Right answer to the heat transfer question. We denote by m' the number of experiments in Spring2013′ for which $K = R$. We find from Table 5 — we simply count the number of entries for which $K = R$ — that $m' = 8$.

Student	Outcome	Student	Outcome	Student	Outcome	Student	Outcome
		2	(N,R)				
5	(N,R)			7	(N,W)		
				11	(N,W)		
13	(N,W)	14	(N,W)	15	(N,W)	16	(N,W)
				19	(N,R)	20	(N,W)
21	(N,R)					24	(N,R)
25	(N,W)			27	(N,W)	28	(N,W)
29	(N,W)			31	(N,W)		
33	(N,W)	34	(N,W)				
		38	(N,W)	39	(N,R)	40	(N,W)
		42	(N,W)			44	(N,W)
45	(N,W)			47	(N,W)		
49	(N,W)					52	(N,R)
		58	(N,R)				
		62	(N,W)				

Table 5: Sample Realization "Spring2013' Dataset." Spring2013' Dataset is the subset of Spring2013 Dataset which includes only those students who have *N*ot yet taken 2.086. Spring2013' Dataset is of size $n' = 30$: only 30 of the students in Spring2013 have *N*ot yet taken 2.086. Spring2013' Dataset retains the student identification number from Spring2013 Dataset of Table 1 to emphasize that Spring2013' Dataset is extracted from Spring2013 Dataset.

Finally, we denote the *conditional frequency* of R *given* N — we write the latter as $R\,|\,N$ — by $\varphi_n(R\,|\,N) \equiv m'/n'$. From Table 5, we conclude that $\varphi_n(R\,|\,N) = 8/30 \approx 0.267$. The interpretation of the conditional frequency should now be clear: $\varphi_n(R\,|\,N)$ is the fraction of the students who have not Taken 2.005 who (also) gave the Right answer to the heat transfer question.

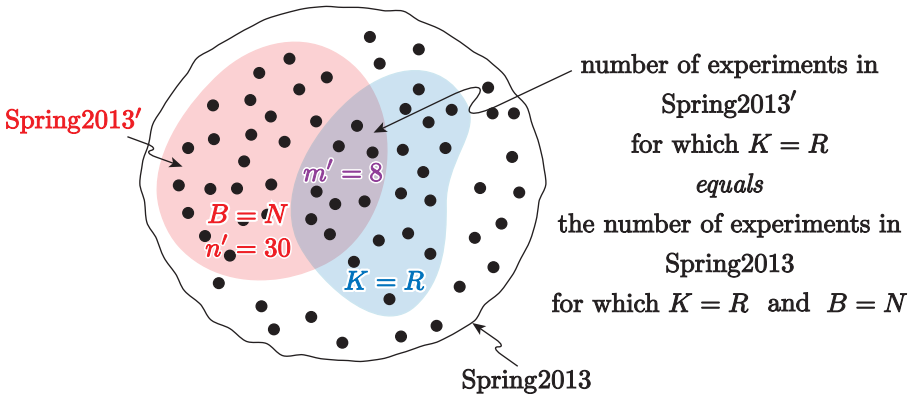


Figure 4: Graphical demonstration: $m' = \#(N, R) = n\varphi_n(N, R)$.

The process above perhaps appears somewhat involved. Fortunately, it is possible to express $\varphi_n(R\,|\,N)$ very simply in terms of the joint and marginal frequencies already developed. We first note that n' is simply $\#(N)$ – the number of entries of Table 1 for which $B = N$ — which we can also write as $\#(N) = n\varphi_n(N)$. We next note that m' is not just the number of experiments in Spring2013′ for which $K = R$, but also the number of experiments in Spring2013 for which $K = R$ and $B = N$ — which is simply $n\varphi_n(N, R)$: we "demonstrate" this equivalence in Figure 4 by the method of dots. We thus obtain the final result $\varphi_n(R\,|\,N) = (m'/n') = (n\varphi_n(N, R))/(n\varphi_n(N))$, or

$$\varphi_n(R\,|\,N) \equiv \frac{\varphi_n(N, R)}{\varphi_n(N)}\ . \tag{19}$$

We know that $\varphi_n(N, R) = 0.125$ and $\varphi_n(N) \approx 0.469$ and hence (19) evaluates to $\varphi_n(R\,|\,N) \approx 0.125/0.469 \approx 0.267$ — in agreement with our results calculated through the intermediary of Spring2013′.

CYAWTP 9. Provide expressions analogous to (19) but now for $\varphi_n(N\,|\,R)$, $\varphi_n(W\,|\,N)$, $\varphi_n(N\,|\,W)$, $\varphi_n(R\,|\,T)$, $\varphi_n(T\,|\,R)$, $\varphi_n(W\,|\,T)$, and $\varphi_n(T\,|\,W)$. Evaluate these conditional frequencies for the the particular sample realization Spring2013 of Table 1. Show that, for any sample realization, $\varphi_n(R\,|\,N) + \varphi_n(W\,|\,N) = 1$, and confirm this result for the particular sample realization Spring2013 of Table 1.

We have obtained $\varphi_n(R\,|\,N) \approx 0.267$ and in **CYAWTP 9** you (should) have obtained $\varphi_{n'}(R\,|\,T) = \varphi_n(T, R)/\varphi_n(T) \approx 0.219/.531 \approx 0.412$. We can now propose some answers to our inference questions. For IQ1, we note that $\varphi_n(R\,|\,N)$ is conspicuously close to the

guessers result of 0.25. For IQ2, we note that $\varphi_n(R \mid T)$ is substantially larger than the guessers result of 0.25.[5] And finally, for IQ3, we note that $\varphi_n(R \mid T)$ is indeed larger, by some considerable margin, than $\varphi_n(R \mid N)$: as a group, the students who have taken 2.005 appear to perform better on the heat transfer question than the group of students who have not taken 2.005.

But what is the interpretation of these results?

> We must ask if the small difference between the guessers result of 0.25 and $\varphi_n(R \,|\, N) \approx 0.267$, and the larger difference between the guessers result of 0.25 and $\varphi_n(R \,|\, T) \approx 0.412$, reflect a true distinction between the two groups — the "haves" and "have nots" — or, rather, intrinsic variations *within* the two groups of students. In other words, are the differences in frequencies we note *statistically significant*, or instead just the result of a likely fluctuation?
>
> On a related note, we must ask if our sample is large enough. As we increase the sample size, n, we attenuate the effect of natural variations with the groups: we reduce the magnitude of likely fluctuations in the frequencies. Certainly we would all agree that we could not draw any conclusions for a sample realization of size $n = 1$, or even $n = 4$ — but is $n = 64$ sufficiently large?
>
> Finally, we must ask if our (rather casual) sampling procedure is adequate. Perhaps "presence at lecture" — recall that our Spring2013 Dataset comprises students present in lecture on a particular day in February — inadvertently selects the more interested students, and misleadingly lifts $\varphi_n(R \,|\, T)$ above the guessers reference of 0.25; had we surveyed *all* students in 2.086 in the Spring of 2013 we might have obtained $\varphi_n(R \,|\, T)$ closer to 0.25. Is our sample realization Spring2013 Datset representative of the larger 2.086 population about which we wish to draw conclusions?

Mathematics alone is not sufficient to address these issues: we must also carefully consider the "implementation" of the mathematics.

But even murkier waters lurk. We (should) know from **CYAWTP 9** that $\varphi_n(T \,|\, R) \approx 0.637$, which is noticeably larger than $\varphi_n(T) \approx 0.531$. We also know that $\varphi_n(R \,|\, T) \approx 0.412$, which is noticeably larger than $\varphi_n(R) \approx 0.344$. But these two quantitatively similar facets of our sample realization do not necessarily share a common interpretation.

> Is it plausible that to provide a *R*ight answer to the heat transfer question *causes* a student to have already *T*aken 2.005? The arrow of time would suggest that no, there is clearly no *causality* — no cause and effect. On the other hand, $\varphi_n(T \,|\, R)$ is higher than $\varphi_n(T)$, and thus there is some *correlation* between the two events: we could exploit the event *R*ight to help us predict the event *T*aken.

[5] Further inspection of the survey results reveals a possible explanation for $\varphi_n(R \,|\, T)$ close to 0.5: a student who has taken 2.005, even if sleepily, can largely rule out the distractors (i) and (ii) in Question 2; if these T students then "guess" between the two remaining options, we arrive at an $R \,|\, T$ frequency of 0.5.

Is it plausible that to have *T*aken 2.005 *causes* a student to provide the *R*ight answer to the heat transfer question? Yes, there is clearly in this case there is correlation but also an argument for *causality*. (If you do not believe there is any causality, you may wish to reconsider your tuition payments.) But we must be careful: perhaps stronger students elect to take 2.005 earlier in the curriculum — hence by the time they take 2.086 — and thus the true reason for the correlation $\varphi_n(R\,|\,T) > \varphi_n(R)$ is not causality between the two events but rather a common underlying factor. (In which case, perhaps you should reconsider your tuition payments.)

In general, solely on the basis of data, we can only identify correlation, not causality. On the other hand, we can often introduce additional variables to refine correlation and better inform our interpretation of the data and hence also the plausibility of causality. For example, in our pedagogical experiment, we might include a third variable, in addition to *B*ackground and *K*nowledge: *G*PA, as a measure of general academic strength or effort. In this way we could perhaps separate the effects of acquired knowledge *versus* natural inclinations.

To close this section, we define conditional frequencies in the most general context. Given any {population, experiment, sample space}, and two events $\mathcal{E}_1$ and $\mathcal{E}_2$, the conditional frequency of $\mathcal{E}_2$ *given* $\mathcal{E}_1$ — we write the latter as $\mathcal{E}_2\,|\,\mathcal{E}_1$ — can be expressed as

$$\varphi_n(\mathcal{E}_2\,|\,\mathcal{E}_1) \equiv \frac{\varphi_n(\mathcal{E}_1 \cap \mathcal{E}_2)}{\varphi_n(\mathcal{E}_1)} \; . \tag{20}$$

The schematic of Figure 4 directly applies to (20) if we interpret the red blob as $\mathcal{E}_1$ and the blue blob as $\mathcal{E}_2$ — the purple blob is hence $\mathcal{E}_1 \cap \mathcal{E}_2$.

CYAWTP 10. Show from the definition (20) that we must obtain for our pedagogical experiment, for *any* sample, (*i*) $\varphi_n(T\,|\,N) \equiv \varphi_n(B = T\,|\,B = N) = 0$, and (*ii*) $\varphi_n(T\,|\,T) \equiv \varphi_n(B = T\,|\,B = T) = 1$.

CYAWTP 11. We can also view the conditional frequency (20) from the outcome perspective: only the outcomes in $\mathcal{E}_1$ shall play a role — a kind of restricted sample space. Express the conditional frequency (20) with reference to the particular sample space and events of Figure 2 in terms of the outcome frequencies, $\varphi_n(O_1), \varphi_n(O_2), \varphi_n(O_3), \varphi_n(O_4)$. Apply the resulting expression to the pedagogical experiment — recall the outcome frequencies are summarized in Table 3 — to evaluate $\varphi_n(R\,|\,N)$.

Bayes' Theorem. We can rewrite our relation (20) to isolate the joint frequency

$$\varphi_n(\mathcal{E}_1 \cap \mathcal{E}_2) = \varphi_n(\mathcal{E}_2\,|\,\mathcal{E}_1)\varphi_n(\mathcal{E}_1) \; ; \tag{21}$$

similarly, exchanging $\mathcal{E}_1$ and $\mathcal{E}_2$, we obtain

$$\varphi_n(\mathcal{E}_2 \cap \mathcal{E}_1) = \varphi_n(\mathcal{E}_1\,|\,\mathcal{E}_2)\varphi_n(\mathcal{E}_2) \; . \tag{22}$$

However, the two events $\mathcal{E}_1 \cap \mathcal{E}_2$ and $\mathcal{E}_2 \cap \mathcal{E}_2$ are equivalent, and thus we may combine our two expressions (21) and (22) to obtain

$$\varphi_n(\mathcal{E}_1\,|\,\mathcal{E}_2) = \frac{\varphi_n(\mathcal{E}_2\,|\,\mathcal{E}_1)\varphi_n(\mathcal{E}_1)}{\varphi_n(\mathcal{E}_2)} \; . \tag{23}$$

The relationship (23) is known as Bayes' Theorem. Bayes' Theorem relates two conditional frequencies in which we exchange the conditioner and conditionee; note that Bayes' Theorem *per se* says nothing about causality.

Bayes' Theorem is valid for any {population, experiment, sample space} and any two events $\mathcal{E}_1$ and $\mathcal{E}_2$ defined over the sample space.

CYAWTP 12. For our pedagogical experiment, and events $\mathcal{E}_1 \equiv B = T$ and $\mathcal{E}_2 \equiv K = R$, Bayes' Theorem tells us that

$$\varphi_n(T \,|\, R) = \frac{\varphi_n(R \,|\, T)\varphi_n(T)}{\varphi_n(R)} \ . \tag{24}$$

Confirm (24) for the particular sample realization Spring2013 Dataset of Table 1.

Independence. We are given a {population, experiment, sample space} and two events $\mathcal{E}_1$ and $\mathcal{E}_2$ defined over the sample space. We say that $\mathcal{E}_1$ and $\mathcal{E}_2$ are *independent* for a particular sample realization if

$$\varphi_n(\mathcal{E}_1 \cap \mathcal{E}_2) = \varphi_n(\mathcal{E}_1)\varphi_n(\mathcal{E}_2) \text{ (\textbf{if} } \mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ are independent for sample realization).} \tag{25}$$

An immediate consequence of (20) and (25) is $\varphi_n(\mathcal{E}_2 \,|\, \mathcal{E}_1) = \varphi_n(\mathcal{E}_2)$: the frequency of $\mathcal{E}_2$ is not affected by whether $\mathcal{E}_1$ happens or not — hence the term independence. Conversely, if two events are not independent — but rather "dependent" — then $\varphi_n(\mathcal{E}_2 \,|\, \mathcal{E}_1) \neq \varphi_n(\mathcal{E}_2)$ and $\varphi_n(\mathcal{E}_1 \,|\, \mathcal{E}_2) \neq \varphi_n(\mathcal{E}_1)$.

In fact, the relation (25) will be rarely satisfied for a particular finite sample. However, in many cases, we can demonstrate or plausibly assume that independence, (25), is indeed valid in the limit that the sample size n tends to infinity (but note that lack of causality is not sufficient argument for independence). In such cases, the relationship (25) will be approximately satisfied,

$$\varphi_n(\mathcal{E}_1 \cap \mathcal{E}_2) \approx \varphi_n(\mathcal{E}_1)\varphi_n(\mathcal{E}_2) \ , \tag{26}$$

for any samples of size n sufficiently large. Approximate independence, (26), in particular in conjunction with (8) or (12), can serve to more easily calculate the frequency of events which can not otherwise be evaluated due to limitations in the data (or in how the data is recorded or presented).

In fact, it is in the limit $n \to \infty$ that independence is most useful, in particular to greatly facilitate the calculation of probabilities, as introduced in the next section.

CYAWTP 13. Reconsider the experiment of **CYAWTP 1** in which we simultaneously flip two coins. Define two events: $\mathcal{E}_1 \equiv$ "coin flipped from the left thumb is a Tail" and $\mathcal{E}_2 \equiv$ "coin flipped from the right thumb is a Tail." Do you expect that $\mathcal{E}_2$ is independent of $\mathcal{E}_1$, at least for larger sample sizes n? Calculate, for your particular sample realization of size $n = 50$, the frequencies $\varphi_n(\mathcal{E}_1)\varphi_n(\mathcal{E}_2)$, $\varphi_n(\mathcal{E}_1)\varphi_n(\mathcal{E}_1)$, and finally $\varphi_n(\mathcal{E}_1 \cap \mathcal{E}_2)$.

CYAWTP 14. Now consider an experiment in which we simultaneously flip *three* coins; you may borrow a friend for the third thumb. Define three events:

$$\mathcal{E}_1 \equiv \text{“coin flip from your left thumb is a Tail”};$$
$$\mathcal{E}_2 \equiv \text{“coin flip from your right thumb is a Tail”};$$
$$\mathcal{E}_3 \equiv \text{“coin flip from your friend's thumb is Tail.”}$$

Based on your empirical evaluation of $\varphi_{n=50}(\mathcal{E}_1)$ *from* ***CYAWTP 13***, estimate — *without any recourse to new (three-coin) data* — the frequency of the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ for a (sufficiently large) simultaneously-flip-three-coins sample realization. Note that the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ corresponds in words to “all three coin flips are Tails.”

5 Probabilities

5.1 From Frequencies to Probabilities

5.1.1 Closure: Pegagogical Experiment

We shall motivate the concept of probability from our pedagogical experiment and the particular data associated with the sample realization Spring2013 Dataset of Table 1. We recall that Spring2013 Dataset is given by $\{H_1 \equiv (B_1, K_1), H_2 \equiv (B_2, K_2), \ldots, H_{n\equiv 64} \equiv (B_{64}, K_{64})\}_{2013}$; for example, from Table 1, we observe that $H_1 \equiv (B_1 = T, K_1 = W)$. In this section, $H_j = (B_j, K_j), 1 \leq j \leq 64$, shall refer to the particular outcomes associated with Spring2013 Dataset.

For any given event $\mathcal{E}$, we now define a cumulative number function $\widetilde{\#}_k(\mathcal{E})$, for $k \leq n$, as the number of occurrences of $\mathcal{E}$ in the set $\{H_1, H_2, \ldots, H_k\}$; we then define our cumulative frequency — a “running” average — as $\widetilde{\varphi}_k(\mathcal{E}) \equiv \widetilde{\#}_k(\mathcal{E})/k$. In words, $\widetilde{\varphi}_k(\mathcal{E})$ is the cumulative frequency of the event $\mathcal{E}$ for a *subsample* of Spring2013 Dataset — in particular, the outcomes associated with *the first k students* of Spring2013 Dataset. Note that $\widetilde{\varphi}_{k=64}(\mathcal{E}) = \varphi_{n\equiv 64}(\mathcal{E})$ since for $\widetilde{\varphi}_{k=64}$ our subsample is in fact the entire sample realization Spring2013 Dataset. We can extend this concept to conditional frequencies: given two events $\mathcal{E}_1$ and $\mathcal{E}_2$, we define the cumulative conditional frequency $\widetilde{\varphi}_k(\mathcal{E}_2 \,|\, \mathcal{E}_1)$ as

$$\widetilde{\varphi}_k(\mathcal{E}_2 \,|\, \mathcal{E}_1) \equiv \widetilde{\varphi}_k(\mathcal{E}_1 \cap \mathcal{E}_2)/\widetilde{\varphi}_k(\mathcal{E}_1) \equiv \widetilde{\#}_k(\mathcal{E}_1 \cap \mathcal{E}_2)/\widetilde{\#}_k(\mathcal{E}_1) \ : \tag{27}$$

we first extract from $\{H_1, H_2, \ldots, H_k\}$ the subsample of experiments for which $\mathcal{E}_1$ occurs; we then calculate the fraction of experiments in this subsample for which $\mathcal{E}_2$ (also) occurs.

What purpose do these cumulative number functions serve? In Figure 5 we present $\widetilde{\varphi}_k(R \,|\, T)$ — arguably our frequency of greatest import from an inference perspective — for $1 \leq k \leq 64$. We can in fact better visualize the result in Figure 6, in which we plot the cumulative conditional frequency only for those values of k for which $B_k = T$: we thus consider only those elements of Spring2013 Dataset which contribute to (and hence change) the cumulative conditional frequency; we thereby eliminate the “plateaus” of Figure 5. We

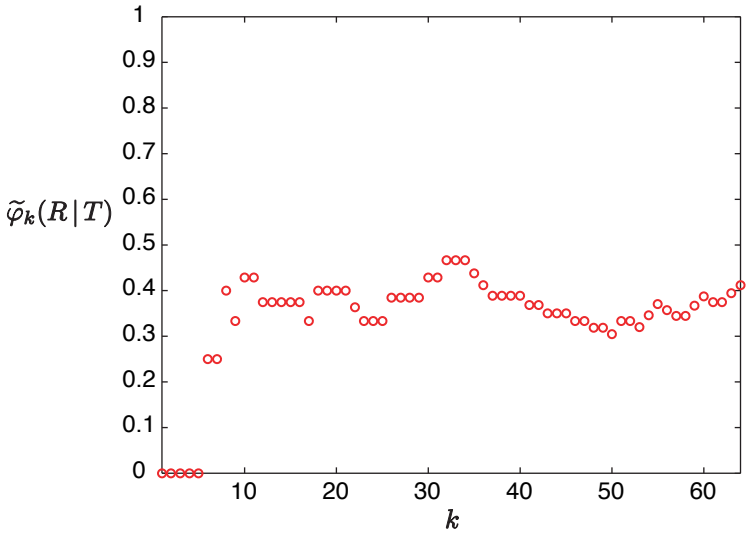


Figure 5: Cumulative conditional frequency $\widetilde{\varphi}_k(R\,|\,T)$ as a function of sample size k for $1 \leq k \leq n \equiv 64$.

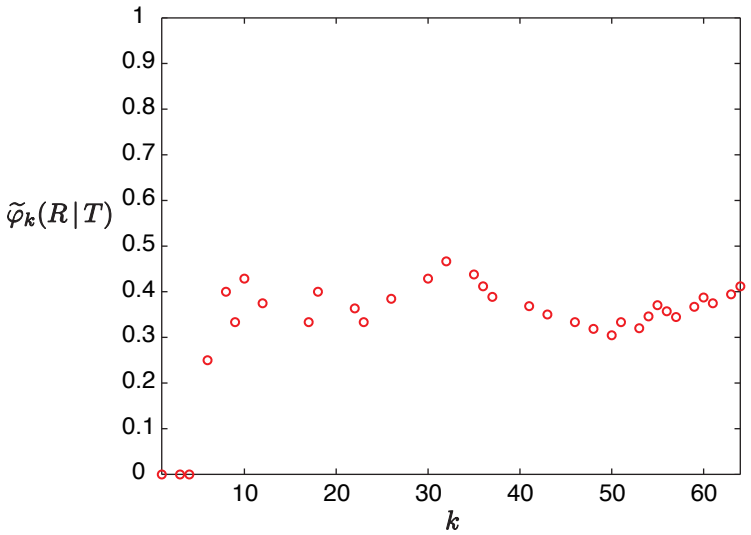


Figure 6: Cumulative conditional frequency $\widetilde{\varphi}_k(R\,|\,T)$ as a function of sample size k for $1 \leq k \leq n \equiv 64$; we plot the result only for k such that $B_k = T$.

observe that $\widetilde{\varphi}_k(R\,|\,T)$ exhibits large oscillations for small k but much smaller fluctuations for larger k. We can indeed imagine from Figure 6 that if we replace Spring2013 Dataset with a much larger sample realization of 2.086 students, say collected over a many-year period, we would observe $\widetilde{\varphi}_k(R\,|\,T)$ approach a limit: $\widetilde{\varphi}_k(R\,|\,T) \to P(R\,|\,T) \equiv p_{R\,|\,T}$ as $k \to \infty$. Here $P(\cdot)$ refers in general to "the probability of," and hence $P(R\,|\,T)$ is read the "the probability of R given T" — which we also provide with the specific label $p_{R\,|\,T}$. In short, we define

probability as the limit of frequency.

On a more pragmatic note, we will of course never enjoy the luxury of an infinite number of 2.086 students. We must thus develop an estimate for $p_{R\,|\,T}$, which we shall denote $\widehat{p}_{R\,|\,T}$. For example, in our current example, we would choose $\widehat{p}_{R\,|\,T} \equiv \widetilde{\varphi}_{k=64}(R\,|\,T) \approx 0.412$. We must now ask in what sense we can estimate the error $\widehat{p}_{R\,|\,T} - p_{R\,|\,T}$ in particular relative to the inference questions described in Section 2.1. We shall focus on IQ2, which we can now pose more precisely:

IQ2 Is $p_{R\,|\,T} > .25$ (the guessers reference)?

(We can also pose IQ2 as a hypothesis we wish to test.) Note our interest in $p_{R\,|\,T}$ derives precisely from the interpretation of probability as frequency in the limit of a large sample — and hence relevant to the entire population.

It would certainly appear from Figure 6 that already for $k = n = 64$ the fluctuations are relatively small. And indeed, it would further appear that the cumulative conditional frequency curve is approaching a limit which is comfortably greater than the group of "guessers" reference of 0.25. But can we improve on these "eyeball" estimates and provide some kind of quantitative assurances that our sample size is indeed sufficiently large? We shall discuss this issue of statistical estimation in a subsequent nutshell. We provide here, solely for pedagogical closure and without explanation, the result, as an in-line appendix:

——— *Appendix* ———

With confidence level (which you can think of, informally, as probability) 95%,

$$\begin{aligned} p_{R\,|\,T} &\geq \hat{p}_{R\,|\,T} - 1.64\sqrt{\frac{\hat{p}_{R\,|\,T}(1-\hat{p}_{R\,|\,T})}{\#(T)}} \\ &\approx 0.412 - 0.138 \approx 0.274 \\ &\geq 0.25 \text{ (the guessers reference)} . \end{aligned} \tag{28}$$

This estimate is known as a one-sided normal-approximation confidence interval.

——— *End Appendix* ———

It would appear that we are safe: students who take 2.005 do indeed perform better than a group of "guessers" on the heat transfer question posed in the Survey. But the safety margin is not generous, and certainly a larger sample size would be welcome, in particular since our error estimate (28) is itself subject to error.

5.1.2 The *Empirical* Approach — and Alternatives

We now generalize the treatment of the previous subsection. We are given a {population, experiment, sample space} and some event $\mathcal{E}$ defined over the sample space. We form a

sample of sample size n in terms of which we can then evaluate the cumulative frequency $\widetilde{\varphi}_k(\mathcal{E}), 1 \leq k \leq n$. We then define $P(\mathcal{E})$, which may also write as $p_{\mathcal{E}}$, as the limit of $\widetilde{\varphi}_k(\mathcal{E})$ as k (and hence n) tends to infinity. Properly speaking, we can only associate probabilities to random experiments. We shall more precisely define "random experiment" in a subsequent nutshell. For our present purposes we simply agree that an experiment (flipping a coin, rolling a die, administering a survey) is random if for any particular realization we can not predict the outcome; we can, of course, predict the frequency of the different outcomes (in our sample space) in the limit of infinitely many realizations — which is precisely our probability. We shall henceforth assume that our experiments are in all cases random experiments.

There are several ways to develop and motivate the notion of probability. In this nutshell we adopt a highly simplified version of the frequentist, or *empirical*, approach espoused by von Mises and further elaborated by Church. However, in some cases, a *mechanistic* approach may be preferred: we posit the outcome probabilities — anticipate the limiting behavior — based on physical arguments; in essence, we propose a "random" model for the physical phenomenon. (In fact, the physical arguments are often quite subtle, implicitly assuming that the phenomenon in question is very sensitive to very small perturbations such that the outcome of an experiment is largely dictated by symmetry or homogeneity.) There are many classic examples in which the mechanistic approach is very successful.

CYAWTP 15. Consider the experiment of flipping a single coin: what would you anticipate for the probability of a Head, $P(H)$? the probability of a Tail, $P(T)$? Now consider the experiment of rolling a single die: what would you anticipate for the probability of rolling a four, $P(4)$?

Most often, probabilities are deduced by a combination of the empirical approach and the mechanistic approach: we posit not the probabilities (of all outcomes) but rather a parametrized distribution of the probabilities over the sample space; we then form a sample realization and conduct experiments to estimate (not the many outcome probabilities but rather) the relatively few parameters which characterize the assumed distribution. We shall study the topic of *statistical parameter estimation* in a subsequent nutshell.

Finally, to close, we introduce another major "school" of probabilistic thought. In some cases, experiments are impossible, very expensive, or dangerous to perform: there is therefore no data, or very little data. Nevertheless, we may wish to understand how the uncertainty in an event may "propagate" to other events. We thus ascribe a *subjective* probability — in this context, a probability is best interpreted as a degree of *belief* — to the event and subsequently apply the laws of probability (to be deduced in the next section, from our rules of frequency). The adjective "Bayesian" is often applied to adherents of this subjective school of probability philosophy. As you can appreciate, the distinction between the "mechanistic" and "subjective" viewpoints can be blurred, and indeed for this reason there is some debate as to whether Bayes himself would be considered a 20^{th}–century Bayesian. (The association of Bayes with subjective probability may also be artifically reinforced by the frequent application of Bayes' Theorem within the subjective context.)

5.2 The Rules of Probability

We now take advantage of the relationship between frequencies and probabilities to deduce the rules of probability. In particular, we may take as our point of departure the results of Section 4.2 for frequencies; we then consider the limit of an arbitrarily large sample, $n \to \infty$, to derive the analogous expressions for probabilities. In practice, the recipe is simple: in all the equations of Section 4 we simply replace $\varphi_n(\mathcal{E})$ with $P(\mathcal{E})$ (or alternatively, $p_{\mathcal{E}}$).

We shall suppose that we are given a population, an experiment, and an associated sample space $\{O_1, O_2, \ldots, O_{n^{\text{outcomes}}}\}$. We further presume that we are provided with outcome probabilities $P(O_i), 1 \leq i \leq n^{\text{outcomes}}$. The outcome probabilities can be deduced from the empirical or mechanistic or even subjective approaches: although we motivate and "derive" the rules of probability from a frequentist perspective, in fact the rules are agnostic to the underlying philosophy.

We first note that, since frequencies lie between 0 and 1, so too must probabilities: $0 \leq P(O_i) \leq 1, 1 \leq i \leq n^{\text{outcomes}}$. We next note from (14) that for any event $\mathcal{E}$, defined as in (13), $\mathcal{E} \equiv \{O_{i_1}, O_{i_2}, \ldots, O_{i_J}\}$,

$$P(\mathcal{E}) = \sum_{i=1}^{J} P(O_{i_j}) \ : \tag{29}$$

the probability of any event may be evaluated as the sum of the probabilities of the outcomes which comprise the event. We further conclude from (17) that

$$\sum_{i=1}^{n^{\text{outcomes}}} P(O_i) = 1 \ . \tag{30}$$

For example, in Figure 2, $0 \leq P(O_i) \leq 1, 1 \leq i \leq n^{\text{outcomes}} \equiv 4$, $P(\mathcal{E}_1)$ is the the sum of $P(O_1)$ and $P(O_2)$, and $P(O_1) + P(O_2) + P(O_3) + P(O_4) = 1$.

We may now proceed to the various relations which relate and combine the frequencies — now also probabilties — of events. In what follows, $\mathcal{E}$ is any event, and $\mathcal{E}_1$ and $\mathcal{E}_2$ are any pair of events. The recipe directly yields

$$P(\mathcal{E}) + P(\bar{\mathcal{E}}) = 1 \ ; \tag{31}$$

$$P(\mathcal{E}_1) + P(\mathcal{E}_2) = 1 \ (\textbf{if } \mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ are mutually exclusive and collectively exhaustive}); \tag{32}$$

$$P(\mathcal{E}_1 \cup \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) - P(\mathcal{E}_1 \cap \mathcal{E}_2) \ ; \tag{33}$$

$$P(\mathcal{E}_1 \cup \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) \ (\textbf{if } \mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ are mutually exclusive}) \ ; \tag{34}$$

$$P(\mathcal{E}_1 \cup \mathcal{E}_2) = 1 - P(\bar{\mathcal{E}}_1 \cap \bar{\mathcal{E}}_2) \ ; \tag{35}$$

$$P(\mathcal{E}_2 \,|\, \mathcal{E}_1) = \frac{P(\mathcal{E}_1 \cap \mathcal{E}_2)}{P(\mathcal{E}_1)} \; ; \tag{36}$$

$$P(\mathcal{E}_1 \cap \mathcal{E}_2) = P(\mathcal{E}_1)\, P(\mathcal{E}_2) \; (\textbf{if } \mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ are independent}) \; . \tag{37}$$

We may also import Bayes' Theorem:

$$P(\mathcal{E}_1 \,|\, \mathcal{E}_2) = \frac{P(\mathcal{E}_2 \,|\, \mathcal{E}_1)\, P(\mathcal{E}_1)}{P(\mathcal{E}_2)} \; . \tag{38}$$

Finally, for M events, $\mathcal{E}_m, 1 \leq m \leq M$,

$$\sum_{m=1}^{M} P(\mathcal{E}_m) = 1 \; (\textbf{if } \mathcal{E}_m, 1 \leq m \leq M, \text{ mutually exclusive and collectively exhaustive}) \; ; \tag{39}$$

$$P(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \cdots \cup \mathcal{E}_M) = 1 - P(\bar{\mathcal{E}}_1 \cap \bar{\mathcal{E}}_2 \cap \cdots \cap \bar{\mathcal{E}}_M) \; ; \tag{40}$$

$$P(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \cdots \cap \mathcal{E}_M) = \prod_{m=1}^{M} P(\mathcal{E}_m) \; (\textbf{if } \mathcal{E}_m, 1 \leq m \leq M \text{ mutually independent}) \; , \tag{41}$$

where $\prod$ refers to the product.

2.086 Numerical Computation for Mechanical Engineers
Fall 2014