Mathematical Statistics
Chapter 1. Probability Theory

1.1 Randomness

The objective of probability is somehow to make sense out of the type of uncertainty known as randomness. In particular, can we use it to make some kind of prediction?

CONCEPT 1.1 **Statistical predictability** means "the proportion of times an event occurs" will converge, over the long run, to a specific value representing the likelihood the event will occur at any given time.

CONCEPT 1.2 **Randomness** is uncertainty that is statistically predictable.

EXAMPLE 1.1 We say that a fair coin is equally likely to to be Heads or Tails – a prediction that favors neither possibility. We also expect that if we flip the coin many times, close to 1/2 of the results will be Heads. If the coin is not fair, we expect the long term proportion of Heads to be some value $p \neq 1/2$. This value would give us the odds $(p : (1-p))$ in favor of Heads.

If we are collecting data for a scientific or business investigation, and our data are random, then we can use statistical predictability to our advantage.

CONCEPT 1.3 Qualified scientific conclusions require repeatable experiments, at least in concept. **Statistical inference** then means making conclusions from random data in such a way that one also can predict the quality (accuracy and reliability) of the conclusions.

EXAMPLE 1.1 (CONT.) Clearly by flipping a coin a large number of times we can estimate the chance of Heads, and the more the flips the better the estimate. In fact, we can state clearly how likely the estimate is to be within, say, Δ of the true value. In other words, statistical predictability gives us both the estimate and an idea of how much the estimate may be in error.

1.2 Sample Spaces and σ-algebras

We see that Concepts 1.1–1.3 implicitly require some kind of experimental context.

DEFINITION 1.4

i. A **random experiment** is a well-defined, repeatable investigation in which exactly one of a set of possible **outcomes** is the experimental result, but just which outcome results is a matter of randomness.

ii. The set of possible outcomes is called the **sample space** and is often denoted $\mathcal{S}$.

iii. An **event** is any subset of $\mathcal{S}$, including $\mathcal{S}$ and $\emptyset$.

EXAMPLE 1.1 (CONT. IN P.1)
Experiment 1: One coin flip. $\mathcal{S} = \{0, 1\}$, where 0 means Tails and 1 means Heads.
Experiment 2: n flips. $\mathcal{S} = \{s = (x_1, \dots, x_n) : x_i \in \{0, 1\}\}$. $\mathcal{S}$ has 2^n outcomes. Possible events include

$$\begin{aligned} A &= \text{“the first two flips are Heads”} \\ &= \{(x_1, \dots, x_n) \in \mathcal{S} : x_1 = x_2 = 1\} \end{aligned}$$

and

$$\begin{aligned} B &= \text{“the total number of Heads is 12”} \\ &= \{(x_1, \dots, x_n) \in \mathcal{S} : x_1 + \dots + x_n = 12\} \end{aligned}$$

Experiment 3: Flip the coin until a Head occurs. $\mathcal{S} = \{1, 2, 3, \dots\}$ is countably infinite.
Experiment 4: An infinite sequence of flips. $\mathcal{S} = \{s = (x_1, x_2, \dots) : x_i \in \{0, 1\}\}$. $\mathcal{S}$ is uncountably infinite. An important event is

$$\begin{aligned} C &= \text{“the limiting proportion of Heads is } 1/2\text{”} \\ &= \{(x_1, x_2, \dots) \in \mathcal{S} : (x_1 + \dots + x_n)/n \to 1/2\}. \end{aligned}$$

Note that the sample space can be infinite, either countably or uncountably.

Events are subsets and adhere to standard set theory. Notation is the usual,

- Union: $A \cup B$, $\cup_{i=1}^n A_i$.
- Intersection: $A \cap B$ or AB, $\cap_{i=1}^n A_i$.
- Complement: A^c.

It is important to note that events are *dynamic* because of the random outcome s. We say “A occurs” to mean the actual experimental outcome is in A or $s \in A$.

Some useful events are

$$\begin{aligned} \bigcup_n A_n &\Leftrightarrow A_n \text{ occurs for some } n \\ \bigcap_n A_n &\Leftrightarrow A_n \text{ occurs for all } n \\ \left(\bigcup_n A_n\right)^c &\Leftrightarrow A_n \text{ does not occur for all } n \\ \left(\bigcap_n A_n\right)^c &\Leftrightarrow \text{some } A_n \text{ does not occur} \end{aligned}$$

DEFINITION 1.5 A collection of events $\mathcal{A}$ is a σ-**algebra** if

i. $\mathcal{A}$ contains $\mathcal{S}$,

ii. $\mathcal{A}$ is closed under complementation: $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$, and

iii. $\mathcal{A}$ is closed under countable unions: $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Basically, $\mathcal{A}$ consists of all "observable" events. It will also be closed under countable intersections. If $\mathcal{S}$ is finite or countable then $\mathcal{A}$ could be all the events, but if $\mathcal{S}$ is uncountable then $\mathcal{A}$ must be restricted some. $\mathcal{A}$ may also be restricted if we only observe partial information.

EXAMPLE 1.1 (CONT. IN P.1)
Experiment 1: Flip a coin once. $\mathcal{S} = \{0, 1\}$ and $\mathcal{A} = \{\emptyset, \{0\}, \{1\}, \mathcal{S}\}$, which is a collection of all possible subsets.
Experiment 2: Flip a coin n times. But suppose we only take note of the number of heads. Then $\mathcal{A}$ consists only of empty set, of events of the form $\{x_1 + \cdots + x_n = k\}, k = 0, 1, \ldots, n$, and of all unions of these events.
Experiment 3: Flip the coin until a Head occurs. $\mathcal{S}$ is countably infinite and $\mathcal{A}$ can be all subsets of $\mathcal{S}$.
Experiment 4: An infinite sequence of flips. $\mathcal{A}$ cannot be all subsets. But we could require it to consist of all events that depend on a finite number of flips, along with all their countable unions, intersections, etc.

EXAMPLE 1.2 Suppose we measure the life time X of a computer under stress. Then we could have $\mathcal{S} = [0, \infty)$. (Unbounded since perhaps we don't really know what the longest lifetime could be.) $\mathcal{A}$ would certainly have to include events of the form $[a, b] =$ "$a \leq X \leq b$", $(a, b) =$ "$a < X < b$", etc., indicating the lifetime falls within a specific interval. The smallest σ-algebra that includes all such events is called the **Borel** σ-**algebra** and it contains everything a statistician would consider.

Events represent the ways we could express the information we observe and it is the events (not the outcomes) that we define probabilities for.

1.3 Axioms and Properties

We can now actually define probability.

DEFINITION 1.6 A **probability space** is $(\mathcal{S}, \mathcal{A}, P)$ where P is a **probability measure** on $\mathcal{A}$ satisfying Kolmogorov's axioms:

i. $P(A) \geq 0$ for all $A \in \mathcal{A}$,

ii. $P(\mathcal{S}) = 1$, and

iii. If $A_1, A_2, \ldots (\in \mathcal{A})$ are disjoint ($A_m \cap A_n = \emptyset$ for $m \neq n$) then $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ (**countable additivity**).

THEOREM 1.7 P is **finitely additive**: if $A_1, \ldots, A_n$ are disjoint then $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.
PROOF First, by considering the sequence $\mathcal{S}, \emptyset, \emptyset, \ldots$, we see that $1 = P(\mathcal{S} \cup (\cup_{n=2}^{\infty} \emptyset)) = P(\mathcal{S}) + \sum_{n=2}^{\infty} P(\emptyset)$. That is, $P(\emptyset)$ must be 0.
Now let $A_{n+1} = A_{n+2} = \cdots = \emptyset$. Then

$$P\left(\cup_{i=1}^{n} A_i\right) = P\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) = \sum_{i=1}^{n} P(A_i).$$

□

Some simple consequences are:

COROLLARY 1.8

i. $P(\emptyset) = 0$,

ii. $P(A^c) = 1 - P(A)$,

iii. $P(A) \leq 1$,

iv. $P(AB) + P(AB^c) = P(A)$,

v. $P(A) + P(B) = P(AB) + P(A \cup B)$,

vi. $A \subset B \Longrightarrow P(A) \leq P(B)$.

PROOF These follow from finite additivity. For example, v. is because

$$P(A) + P(B) = \{P(AB) + P(AB^c)\} + \{P(AB) + P(A^cB)\}$$

and

$$P(AB) + P(AUB) = P(AB) + \{P(AB) + P(AB^c) + P(A^cB)\}.$$

□

THEOREM 1.9

i. (Boole) $P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$.

ii. (Bonferroni) If $P(A_i) \geq 1 - a_i$ then $P(\cap_{i=1}^{n} A_i) \geq 1 - \sum_{i=1}^{n} a_i$.

PROOF i. Repeatedly apply Cor. 1.8.v:

$$P(\cup_{i=1}^{n} A_i) \leq P(A_1) + P(\cup_{i=2}^{n} A_i) \leq \cdots \leq \sum_{i=1}^{n} P(A_i).$$

ii. Apply DeMorgan's rule, Cor. 1.8.ii and Boole's inequality:

$$P(\cap_{i=1}^n A_i) = 1 - P(\cup_{i=1}^n A_i^c) \geq 1 - \sum_{i=1}^n P(A_i^c) \geq 1 - \sum_{i=1}^n a_i.$$

□

EXAMPLE 1.3 Suppose three statistical methods each of which has more than 95% chance of providing an accurate result (95% confidence). Then the chance all three are accurate is at least 85%.

EXAMPLE 1.1 (CONT. IN P.1) Flip a fair coin 5 times. $\mathcal{S}$ has $2^5 = 32$ equally likely outcomes. The chance of at least one Head is

$$P(x_1 + \cdots + x_5 > 0) = 1 - P(\{(0,0,0,0,0)\}) = 1 - \tfrac{1}{32} = \tfrac{31}{32}.$$

THEOREM 1.10 Suppose $\mathcal{S}$ is either finite or countable and $A \subset \mathcal{S}$. Then

$$P(A) = \sum_{n:s_n \in A} P(\{s_n\}).$$

PROOF By countable additivity, since $A = \cup_{n:s_n \in A}\{s_n\}$ which is a union of disjoint events. □

EXAMPLE 1.4 Flip a coin until Heads and observe only the number of flips. $\mathcal{S} = \{1, 2, \ldots\}$. The chance the number of flips is even would be

$$P(\text{“even number of flips”}) = P(\{2\}) + P(\{4\}) + \cdots = 1/3.$$

If $\mathcal{S}$ is countable we need only to know the probability of each outcome. But this is *not* true if $\mathcal{S}$ is uncountable.

Another frequently used result:

THEOREM 1.11 Let $B_1, B_2, \ldots$ be a **partition** for $\mathcal{S}$: $B_1, B_2, \ldots$ are disjoint and their union is $\mathcal{S}$. Then, for any event A, $P(A) = \sum_n P(AB_n)$.
PROOF $A = \cup_n AB_n$ and $AB_1, AB_2, \ldots$ are also disjoint, so it follows by countable (or finite) additivity. □

EXAMPLE 1.5 Roll a die until a 6 appears and stop. What is the probability that a 1 never appears?
SOLUTION Let A = “no 1's” and B_n = “number of rolls is n”. The B_n's partition $\mathcal{S}$. Since AB_n = “$n-1$ rolls of 2, 3, 4 or 5 followed by one roll of 6”, we get

$$P(AB_n) = \frac{4 \times 4 \times \cdots \times 4 \times 1}{6^n} = \frac{1}{6}\left(\frac{2}{3}\right)^{n-1},$$

and thus

$$P(A) = \sum_{n=1}^{\infty} \frac{1}{6}\left(\frac{2}{3}\right)^{n-1} = \frac{1}{6}\left(\frac{1}{1-2/3}\right) = \frac{1}{2}.$$

(Intuition says a 1 is not more likely to appear before a 6 than a 6 is to appear before a 1. But this thought is not a proof.)

We will see that partitioning one event (or value) according to another frequently reduces the work in a calculation. It is also very useful to compute probabilities as limits.

THEOREM 1.12 (Continuity)

i. Suppose $A_1 \subset A_2 \subset \cdots$ (**increasing sequence of events**) and $A = \cup_{n=1}^{\infty} A_n$. Then $P(A) = \lim_{n\to\infty} P(A_n)$.

ii. Suppose $A_1 \supset A_2 \supset \cdots$ (**decreasing sequence of events**) and $A = \cap_{n=1}^{\infty} A_n$. Then $P(A) = \lim_{n\to\infty} P(A_n)$.

PROOF

i. Let $B_1 = A_1$ and $B_n = A_n \cap (\cup_{i=1}^{n-1} A_i)^c$ for $n > 1$. Note that the B_n's are disjoint and $A_n = \cup_{i=1}^{n} B_i$. Thus

$$\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} P\left(\cup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) = \sum_{i=1}^{\infty} P(B_i).$$

But $s \in$ some $A_n \iff s \in$ some B_n so $A = \cup_{i=1}^{\infty} B_i$ and

$$P(A) = P\left(\cup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i).$$

ii. (exercise).

□

EXAMPLE 1.5 (CONT. IN P.5) What is the probability a 6 never appears?
SOLUTION Let $A_n =$ "no 6 in the first n rolls". Then $A_1 \supset A_2 \supset \cdots$, $A = \cap_{n=1}^{\infty} A_n =$ "no 6's ever" and $P(A) = \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} 5^n/6^n = 0$. (But A is not the empty set.)

1.4 Counting Rules

Classical probability and the ideas of random sampling depend on being able to count the number of ways a selection can be made.

THEOREM 1.13 (Counting Selections)

i. (Product Rule) If selection i has n_i possibilities, $i = 1, \ldots, k$, irrespective of the other selections, then the total number of possibilities is $\prod_{i=1}^{k} n_i$.

ii. (Ordering Rule) There are $n!$ ways to order n items.

iii. (Permutation Rule) The number of ordered (order identified) selections of n items from N is $\frac{N!}{(N-n)!}$.

iv. (Combinations Rule) The number of subsets (unordered selections) of n items from N is $\binom{N}{n} = \frac{N!}{n!(N-n)!}$.

EXAMPLE 1.6 (Quality Inspection) Imagine a production lot of N computers from which we take a random sample of $n \ll N$ to inspect. The lot has an (unknown) defective rate of $p = M/N$. If we observe X defectives in the sample, we can estimate p with the sample rate, $\hat{p} = X/n$. What can we predict about the values of this estimate (or, equivalently, of X)?
SOLUTION The sample space consists of $\binom{N}{n}$ equally likely possible samples (ignoring order). This is **Sampling Without Replacement** (SWOR).
As there are $\binom{M}{x}$ ways to choose x defectives from among the M defectives, and $\binom{N-M}{n-x}$ ways to select the non-defectives,

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad 0 \le x \le n,\ n - N + M \le x \le M.$$

To determine when this probability is largest, note that

$$\begin{aligned} \tfrac{P(X=x-1)}{P(X=x)} &= \tfrac{x!(M-x)!(n-x)!(N-M-n+x)!}{(x-1)!(M-x+1)!(n-x+1)!(N-M-n+x-1)!} \\ &= \tfrac{x(N-M-n+x)}{(M-x+1)(n-x+1)} \\ &= 1 + \tfrac{(N+2)x-(M+1)(n+1)}{(M-x+1)(n-x+1)}. \end{aligned}$$

Thus $P(X = x) > P(X = x-1) \iff x < \frac{(M+1)(n+1)}{N+2}$ and the most likely value for $\hat{p}$ is $\frac{1}{n}\lfloor\frac{(M+1)(n+1)}{N+2}\rfloor \doteq \frac{M}{N} = p$.

EXAMPLE 1.7 (Independent Polling) A pollster calls n voters, selecting each at random from the entire population of N voters, irrespective of previous selections. Suppose $p = M/N$ is the proportion in the population that would answer Yes to the pollster. Let X be the number of Yes responses in the sample. Again, the pollster would estimate p with $\hat{p} = X/n$. How do we predict X now?
SOLUTION By the product rule there are N^n equally likely possible samples (ordered). This is **Sampling With Replacement** (SWR). The number of samples for which there are x Yes responses (and $n - x$ No's) is

$$M^x \times (N-M)^{n-x} \times \text{\# ways to place } x \text{ Yes's among } n \text{ responses.}$$

Thus,

$$P(X = x) = \frac{\binom{n}{x}M^x(N-M)^{n-x}}{N^n} = \binom{n}{x}p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

The pollster hopes to estimate p to within a small value Δ. The chance of this is

$$P\{n(p-\Delta) \le X \le n(p+\Delta)\} = \sum_{n(p-\Delta)\le x\le n(p+\Delta)} \binom{n}{x}p^x(1-p)^{n-x}.$$

For example, if $n = 1000$, $p = .4$ and $\Delta = .03$ then

$$P(370 \leq X \leq 430) = \sum_{370 \leq x \leq 430} \binom{n}{x} p^x (1-p)^{n-x} = .95108.$$

1.5 Conditional Probability and Bayes' Theorem

We come now to our first look at one of the most important ideas of probability.

EXAMPLE 1.8 (Medical Testing) A hidden disease D inhabits 40% of the population. A medical test is available but it has a 5% false positive rate and a 20% false negative rate. Just what does this mean? Let us the population size be N, so .4N individuals are diseased. Of those, 20% or .08N will fail to show a positive response "P" to the test – they have a false negative response. Thus $P(\text{"D and not P"}) = \frac{.08N}{N} = .08$. Note that this rate is relative to the entire population, while the false negative rate of 20% is relative only to the diseased subpopulation. To interpret the 20% we essentially focus only on those who are diseased and treat them as a given whole. Note, also that the calculation does not actually depend on the value of N: $.20 = \frac{.08}{.40}$.

DEFINITION 1.14 Let A and B be events. The **conditional probability** of A, given B, is $P(A|B) = \frac{P(AB)}{P(B)}$, if $P(B) > 0$.

EXAMPLE 1.8 (CONT.) So the false negative rate (20%) is $P(P^c|D)$, the conditional probability of "not positive", given "diseased". Likewise, the false positive rate is the conditional probability of "positive", given "not diseased":

$$.05 = P(P|D^c) = \frac{P(P \cap D^c)}{P(D^c)}.$$

We can then calculate $P(P \cap D^c) = .05(1 - .40) = .03$.

THEOREM 1.15 (Multiplication Rule)

i. $P(AB) = P(B)P(A|B)$ (with the obvious interpretation $P(AB) = 0$ if $P(B) = 0$).

ii. $P(A_1 A_2 \cdots A_n) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_1 \cdots A_{n-1})$.

EXAMPLE 1.8 (CONT.) Suppose the actual disease rate is unknown, but doctors are confident of the false negative and false positive rates and they know that 35% of the population tests positive. We can then "condition" on disease status:

$$\begin{aligned} .35 &= P(P) = P(PD) + P(PD^c) = P(D)P(P|D) + P(D^c)P(P|D^c) \\ &= (1 - .2)P(D) + .05(1 - P(D)). \end{aligned}$$

Hence $P(D) = .40$.

COROLLARY 1.16 Assume $P(B) > 0$.

i. $P(B|B) = 1$.

ii. $A \subset B \Longrightarrow P(A|B) = \frac{P(A)}{P(B)}$.

iii. $AB = \emptyset \Longrightarrow P(A|B) = 0$.

THEOREM 1.17 Suppose $B_1, B_2, \ldots$ is a partition of $\mathcal{S}$.

i. (Total Probability) $P(A) = \sum_n P(A|B_n)P(B_n)$.

ii. (Bayes' Rule) $P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_n P(A|B_n)P(B_n)}$ for each i.

PROOF

i. Use the partitioning rule (Thm. 1.11) and the multiplication rule (Thm. 1.15.i).

ii. Use the law of total probability (Thm 1.17.i) and the definition of conditional probability (Def. 1.14).

□

It is easier to remember the derivation of Bayes' rule than to remember the formula itself.

EXAMPLE 1.9 (Two-State Markov Chain) Suppose the signal a radio receives has two states: 1 = "Clear", meaning a clear signal is received, and 0 = "Fade", meaning the signal is too weak. Let X_i be the value of the state at time i, where $i \geq 0$. Suppose α is the chance the state changes from 1 to 0, and β is the chance it changes from 0 to 1. Let $p_i = P(X_i = 1)$. Then

$$\begin{aligned} p_i &= P(X_i = 1|X_{i-1} = 0)P(X_{i-1} = 0) + P(X_i = 1|X_{i-1} = 1)P(X_{i-1} = 1) \\ &= \beta(1 - p_{i-1}) + (1 - \alpha)p_{i-1}. \end{aligned}$$

This gives a recursive way to calculate the probabilities.
If the process is in equilibrium then $p_i = p_{i-1}$ and we get $p_i = \frac{\beta}{\alpha+\beta}$ for all i.
Suppose the process is "Clear" at time i. What is the chance it also was "Clear" at time $i - 1$? By Bayes' rule,

$$P(X_{i-1} = 1|X_i = 1) = \frac{(1-\alpha)p_{i-1}}{\beta(1 - p_{i-1}) + (1 - \alpha)p_{i-1}}.$$

In equilibrium, this is $1 - \alpha$.

EXAMPLE 1.10 Suppose 60% of marriages have children, 40% of marriages end in divorce and 20% of marriages with children end in divorce. So does $24\% = .40 \times 60\%$ of marriages divorce with children? The answer is no, because only 20% of marriages with children end in divorce. The correct value is $.20 \times 60\% = 12\%$. On the other hand, $70\% = (.40 - .12)/(1 - .60)$ of marriages without children end in divorce.

Example 1.11 You roll two fair dice hoping for a total of 7 (probability 1/6). After the roll one die is hidden, but you see that the other is a 4. What is the chance you have a total of 7?
Solution If you had rolled just one die and seen it was a 4, then to get a 7 you would have to roll a 3 with the second die; the chance again is 1/6. But in this case you don't know which die is hidden and which you see. There are 11 ways that one die could be a 4, two of which have the other die a 3. So the updated chance "total is 7" is 2/11.

Theorem 1.18 Suppose $P(B) > 0$ and define $P^B(A) = P(A|B)$. Then P^B is a valid probability measure.
Proof P^B satisfies Kolmogorov's axioms (Def. 1.6). In particular, if $A_1, A_2, \ldots$ are disjoint then

$$P^B(\cup_n A_n) = \frac{P(\cup_u A_n B)}{P(B)} = \frac{\sum_n P(A_n B)}{P(B)} = \sum_n P^B(A_n).$$

□

In other words, if B is fixed we compute with $P(A|B)$ as we would with any probability: $P(A^c|B) = 1 - P(A|B)$, etc. But, as in Ex. 1.11, the rules of probability only work right if we are talking about the same condition B throughout.

Example 1.5 (cont. in p.5) Roll a die until a 6 appears. Given that a 6 appears on the $(n+1)$th roll, what is the distribution (of probabilities) for the number of 1's that have appeared?
Solution On the condition a 6 first appears on roll $n+1$, we have an experiment for which there are n selections with replacement from $\{1,2,3,4,5\}$ ($N = 5, M = 1$). We can calculate as if this defines the sample space. So by Ex. 1.7 (p.7) we can state

$$P(\text{“1 appears } x \text{ times”} \,|\, \text{“6 appears first on } (n+1)\text{th roll”}) = \binom{n}{x}(.2)^x(.8)^{n-x}, \quad x = 0, 1, \ldots, n.$$

1.6 Independence

If investigating the dependence between events or variables is the essence of statistics, then modeling them by starting from simple, independent terms is a major objective of probability.

Definition 1.19 Suppose $0 < P(B) < 1$.

i. A and B are **dependent** if $P(A|B) \neq P(A|B^c)$.

ii. A and B are **independent** if $P(A|B) = P(A|B^c)$.

THEOREM 1.20 A and B are independent $\iff P(AB) = P(A)P(B) \iff P(A|B) = P(A)$.
PROOF We have A and B are independent if and only if

$$\begin{aligned} & \frac{P(AB)}{P(B)} = \frac{P(AB^c)}{P(B^c)} = \frac{P(A) - P(AB)}{1 - P(B)} \\ \iff \quad & P(AB) = P(A)P(B) \\ \iff \quad & P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A). \end{aligned}$$

□

Thm. 1.20 gives a definition for independence that is useful even if $P(B) = 0$ or $P(B) = 1$. In those cases B is independent of every event, including itself.

EXAMPLE 1.11 (CONT. IN P.10) Consider rolling 2 dice. Let $A_i =$ "1st die is i" and $B_j =$ "2nd die is j". The sample space consists of 36 equally likely outcomes (i, j), $i, j = 1, \ldots, 6$. So

$$P(A_i) = \frac{6}{36} = \frac{1}{6}, \quad P(B_j) = \frac{6}{36} = \frac{1}{6}$$

and $P(A_iB_j) = 1/36$. Since $P(A_iB_j) = P(A_i)P(B_j)$, we see A_i and B_j are independent. Since this is true for any i and j, we can say the two dice are independent.

EXAMPLE 1.7 (CONT. IN P.7) Polling with replacement. Suppose there are N voters in the population, M of whom would respond Yes. If you sample with replacement the two responses are independent. (exercise).

Independence is a specialized relationship and not to be assumed lightly.

DEFINITION 1.21 Events $A_1, A_2, \ldots, A_n$ are **mutually independent** if for every subcollection $A_{i_1}, \ldots, A_{i_k}$,

$$P(A_{i_1}, \ldots, A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j}).$$

The definition says this equality must hold for *any* subcollection, not just for all n events, and not just for all pairs of events (which would be **pairwise independence**). However, the definition also implies that each subcollection is mutually independent.

EXAMPLE 1.7 (CONT. IN P.7) Sample n voters with replacement. The n responses are independent. (exercise).

Clearly, knowing if A occurs is equivalent to knowing if A^c occurs. Thus, independence of events should encompass their complements.

THEOREM 1.22 Suppose $A_1, \ldots, A_n$ are mutually independent events. Let $A_{i_1}, \ldots, A_{i_k}$ and $A_{j_1}, \ldots, A_{j_m}$ be two subcollections such that $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_m\}$ are disjoint. Then $A_{i_1}, \ldots, A_{i_k}, A^c_{j_1}, \ldots, A^c_{j_m}$ are independent and

$$P(A_{i_1} \cdots A_{i_k} \cap A^c_{j_1} \cdots A^c_{j_m}) = \prod_{s=1}^{k} P(A_{i_s}) \prod_{t=1}^{m} \{1 - P(A_{jt})\},$$

with the obvious interpretation if $k = 0$ or $m = 0$.

PROOF For independent events $B_1, \ldots, B_m$,

$$P(B_1 \cdot B_{m-1} B^c_m) = P(B_1 \cdots B_{m-1}) - P(B_1 \cdots B_{m-1} B_m)$$

$$= \prod_{i=1}^{m-1} P(B_i) - \prod_{i=1}^{m} P(B_i) = \prod_{i=1}^{m-1} P(B_i) \times \{1 - P(B_m)\}.$$

$B_1, \ldots, B_m$ could be any subcollection of independent events so we can conclude, for example, that switching A_{j_1} to $A^c_{j_1}$ retains the independence. We just apply this iteratively, attaching another A^c_{jt} at each step. □

EXAMPLE 1.12 Suppose $A_1, \ldots, A_n$ are independent events, each with probability p. Let $B_k =$ "exactly k of the A_i's occur". Given a particular choice of which A_i's occur (and which do not), Thm. 1.22 indicates a probability of $p^k(1-p)^{n-k}$. There are $\binom{n}{k}$ ways to select k of the n events. Therefore,

$$P(B_k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Compare this with the probability that exactly k of n voters (sampled with replacement) will respond Yes (Ex. 1.7). Here, however, there is no sampling frame and we did not use counting to get the probability.

EXAMPLE 1.13 Consider flipping an unfair coin with $p = P(\text{Head on } n\text{th flip})$, $0 < p < 1$. We cannot compute by counting now. Assume the flips are independent. How long until a Head is observed? until the kth Head is observed?

SOLUTION If the first Head is observed on flip n, there must be $n - 1$ Tails beforehand. By the independence,

$$P(\text{"1st Head is on } n\text{th flip"}) = p(1-p)^{n-1}, \quad n = 1, 2, \ldots$$

If the kth Head is on flip n, there are $k - 1$ other Heads placed among the previous $n - 1$ flips. There are $\binom{n-1}{k-1}$ ways to do this, so

$$P(\text{"}k\text{th Head is on } n\text{th flip"}) = \binom{n-1}{k-1} p^k (1-p)^{n-k}, \quad n = k, k+1, k+2, \ldots$$