Laws of Probability, Bayes' theorem, and the Central Limit Theorem

2022 Penn State Astrostatistics Summer School

David Hunter
Department of Statistics
Penn State University

Adapted from notes prepared by
Rahul Roy and RL Karandikar,
Indian Statistical Institute, Delhi

Seventeenth Penn State Astrostatistics Summer School
June, 2022

Prologue: Blindsight & insight in visuo-spatial neglect

This is a 1988 paper by Marshall and Halligan in *Nature*.

> *...The patient, P.S., had sustained right cerebral damage.... P.S. was presented simultaneously with two line drawings of a house, in one of which the left side was on fire. She judged that the drawings were identical; yet when asked to select which house she would prefer to live in, she reliably chose the house that was not burning.*

Actually, she chose the non-burning house 14 times in 17 trials.

I heard about this paper from Rob Kass of CMU, who writes

> *I love this because if it were 17 out of 17, we're done. If it were 9 out of 17, that's chance. But 14 out of 17? That sounds pretty convincing, but I want to do a calculation.*

Prologue: Blindsight & insight in visuo-spatial neglect

This is a 1988 paper by Marshall and Halligan in *Nature*.

> *...The patient, P.S., had sustained right cerebral damage.... P.S. was presented simultaneously with two line drawings of a house, in one of which the left side was on fire. She judged that the drawings were identical; yet when asked to select which house she would prefer to live in, she reliably chose the house that was not burning.*

Actually, she chose the non-burning house 14 times in 17 trials.

I heard about this paper from Rob Kass of CMU, who writes

> *I love this because if it were 17 out of 17, we're done. If it were 9 out of 17, that's chance. But 14 out of 17? That sounds pretty convincing, but I want to do a calculation.*

Prologue: Blindsight & insight in visuo-spatial neglect

This is a 1988 paper by Marshall and Halligan in *Nature*.

> *...The patient, P.S., had sustained right cerebral damage.... P.S. was presented simultaneously with two line drawings of a house, in one of which the left side was on fire. She judged that the drawings were identical; yet when asked to select which house she would prefer to live in, she reliably chose the house that was not burning.*

Actually, she chose the non-burning house 14 times in 17 trials.

I heard about this paper from Rob Kass of CMU, who writes

> *I love this because if it were 17 out of 17, we're done. If it were 9 out of 17, that's chance. But 14 out of 17? That sounds pretty convincing, but I want to do a calculation.*

Rob Kass Continues:

> *When data are collected repeatedly under conditions that are as nearly identical as an investigator can make them, the measured responses nevertheless exhibit variation. The most fundamental principle of the statistical paradigm, its starting point, is that variation may be described by probability. ... Probability describes familiar games of chance such as rolling dice, so when we use probability to describe variation, we're making an analogy. We do not know all the reasons why one measurement is different than another, so it's as if the variation in the data were generated by a gambling device.*

— *Analysis of Neural Data* by Kass, Eden, and Brown (2014)

Outline

Outline

Do Random phenomena exist in Nature?

- Is a coin flip random?
  Not really, given enough information. But modeling the outcome as random gives a parsimonious representation of reality.

Randomness is not total unpredictability; we may quantify that which is unpredictable.

Do Random phenomena exist in Nature?

If Mathematics and Probability theory were as well understood several centuries ago as they are today but the planetary motion was not understood, perhaps people would have modeled the occurrence of a Solar eclipse as a random event and assigned a daily probability based on past empirical evidence.

Subsequently, someone would have revised the model, observing that a solar eclipse occurs only on a new moon day. After more time, the phenomenon would be completely understood and the model changed from a stochastic, or random, model to a deterministic one.

Do Random phenomena exist in Nature?

Thus, we often come across events or measurements whose outcome is (partly) uncertain. The uncertainty could be because of

- our inability to observe accurately all the inputs required to compute the outcome;
- excessive cost of observing all the inputs;
- lack of understanding of the phenomenon;
- dependence on choices to be made in the future, like the outcome of an election.

Probability gives us a way to model uncertainty, i.e., to provide a mathematical structure to the notion of uncertainty.

Outline

Existing Intuition is Useful

The structure needed to understand a coin toss is intuitive. We assign a probability $1/2$ to the outcome HEAD and a probability $1/2$ to the outcome TAIL of appearing.

Similarly, for each of the outcomes 1,2,3,4,5,6 of the throw of a die, we assign a probability $1/6$ of appearing.

Similarly, for each of the outcomes $000001, \ldots, 999999$ of a lottery ticket, we assign a probability $1/999999$ of being the winning ticket.

Sample space: Set of possible outcomes

More generally, associated with any experiment we have a sample space Ω consisting of outcomes $\{o_1, o_2, \ldots, o_m\}$.

- Coin Toss: Ω = {H, T}
- One die: Ω = {1, 2, 3, 4, 5, 6}
- Lottery: Ω = {1, . . ., 999999}

Each outcome is assigned a probability according to the physical understanding of the experiment.

- Coin Toss: $p_H = 1/2$, $p_T = 1/2$
- One die: $p_i = 1/6$ for $i = 1, \ldots, 6$
- Lottery: $p_i = 1/999999$ for $i = 1, \ldots, 999999$

Note that in each example, the sample space is *finite* and the probability assignment is *uniform* (i.e., the same for every outcome in the sample space), but this need not be the case.

Discrete Sample Space: Countably many outcomes

- More generally, for an experiment with a *finite* sample space $\Omega = \{o_1, o_2, \ldots, o_m\}$, we assign a probability p_i to the outcome o_i for every i in such a way that the probabilities add up to 1, i.e., $p_1 + \cdots + p_m = 1$.
- In fact, the same holds for an experiment with a *countably infinite* sample space. (Example: Roll one die until you get your first six.)
- A finite or countably infinite sample space is sometimes called *discrete*.
- Uncountably infinite sample spaces exist, and there are some additional technical issues associated with them. We will hint at these issues without discussing them in detail.

An event is a subset of the sample space (almost)

- A subset $E \subseteq \Omega$ is called an *event*.
- For a discrete sample space, this may if desired be taken as the mathematical definition of *event*.

Technical word of warning: If Ω is uncountably infinite, then we cannot in general allow arbitrary subsets to be called events; in strict mathematical terms, a *probability space* consists of:

- a sample space Ω,
- a set $\mathcal{F}$ of subsets of Ω that will be called the events,
- a function P that assigns a probability to each event and that must obey certain axioms.

Often, we don't have to worry about these technical details in practice.

Random Variables link the sample space with the reals

Back to the dice. Suppose we are gambling with one die and have a situation like this:

outcome	1	2	3	4	5	6
net dollars earned	-8	2	0	4	-2	4

Our interest in the outcome is only through its association with the monetary amount. So we are interested in a function from the outcome space Ω to the real numbers $\mathbb{R}$. Such a function is called a *random variable*.

Technical word of warning: A random variable must be a *measurable* function from Ω to $\mathbb{R}$, i.e., the inverse function applied to any interval subset of $\mathbb{R}$ must be an event in $\mathcal{F}$. For our discrete sample space, *any* map $X : \Omega \to \mathbb{R}$ works.

Random Variables may define events

outcome	1	2	3	4	5	6
net dollars earned	-8	2	0	4	-2	4

- Let X be the amount of money won on one throw of a die. We are interested in events of the form $\{\omega \in \Omega : X(\omega) = x\}$ for a given x. (Shorthand notation: $\{X = x\}$)

$$\begin{array}{ll} \text{for } x = 0, & \{X = x\} \text{ is the event } \{3\}. \\ \text{for } x = 4, & \{X = x\} \text{ is the event } \{4, 6\}. \\ \text{for } x = 10, & \{X = x\} \text{ is the event } \emptyset. \end{array}$$

- Notation is informative: Capital "X" is the random variable whereas lowercase "x" is some fixed value attainable by the random variable X.
- Thus X, Y, Z might stand for random variables, while x, y, z could denote specific points in the ranges of X, Y, and Z, respectively.

Random Variables are a way to quantify uncertainty

outcome	1	2	3	4	5	6
net dollars earned	-8	2	0	4	-2	4

- The probabilistic properties of a random variable are determined by the probabilities assigned to the outcomes of the underlying sample space.
- Example: To find the probability that you win 4 dollars, i.e. $P(\{X = 4\})$, you want to find the probability assigned to the event $\{4, 6\}$. Thus

$$P\{\omega \in \Omega : X(\omega) = 4\} = P(\{4, 6\}) = (1/6) + (1/6) = 1/3.$$

Disjoint events: Probability of union = sum of probabilities

$$P\{\omega \in \Omega : X(\omega) = 4\} = P(\{4, 6\}) = (1/6) + (1/6) = 1/3.$$

- Adding $1/6 + 1/6$ to find $P(\{4, 6\})$ uses a probability axiom known as *finite additivity*:

  Given disjoint events A and B, $P(A \cup B) = P(A) + P(B)$.

- In fact, any probability measure must satisfy *countable additivity*:

  Given mutually disjoint events $A_1, A_2, \ldots$, the probability of the (countably infinite) union equals the sum of the probabilities.

  N.B.: "Disjoint" means "having empty intersection".

Probability Mass Function: Summary of a R.V.

$$P\{\omega \in \Omega : X(\omega) = 4\} = P(\{4, 6\}) = (1/6) + (1/6) = 1/3.$$

- $\{X = x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\}$
- If we summarize the possible nonzero values of $P(\{X = x\})$, we obtain a function of x called the *probability mass function* of X, sometimes denoted $f(x)$ or $p(x)$ or $f_X(x)$:

$$f_X(x) = P(\{X = x\}) = \begin{cases} 1/6 & \text{for } x = 0 \\ 1/6 & \text{for } x = 2 \\ 2/6 & \text{for } x = 4 \\ 1/6 & \text{for } x = -2 \\ 1/6 & \text{for } x = -8 \\ 0 & \text{for any other value of } x. \end{cases}$$

Probability functions must satisfy three axoims

Any probability function P must satisfy these three axioms, where A and A_i denote arbitrary events:

- $P(A) \geq 0$ *(Nonnegativity)*
- If A is the whole sample space Ω then $P(A) = 1$
- If $A_1, A_2, \ldots$ are mutually exclusive (i.e., disjoint, which means that $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \qquad \textit{(Countable additivity)}$$

Technical digression: If Ω is uncountably infinite, it is impossible to define P satisfying these axioms if “event” means “any subset of Ω.”

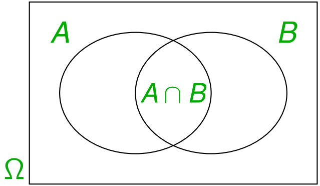
Inclusion-Exclusion Rule: Probability of a union

- Countable additivity requires that if A, B are disjoint,

$$P(A \cup B) = P(A) + P(B).$$

- More generally, for *any* two events A and B,

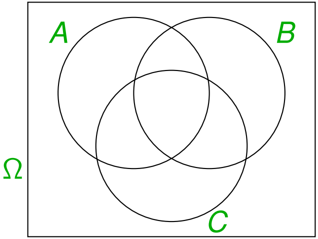
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$



Inclusion-Exclusion Rule: Probability of a union

- Similarly, for three events A, B, C:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C) \end{aligned}$$



- This identity has a generalization to n events called the *inclusion-exclusion rule*.

Probabilities can be based on counting outcomes

- Simplest case: Due to inherent symmetries, we can model each outcome in Ω as being equally likely.
- When Ω has m equally likely outcomes $o_1, o_2, \ldots, o_m$,

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{m}.$$

- **Well-known example (not easy):** If n people shuffle their hats, what is the probability that at least one person gets their own hat back?
  The answer,

$$1 - \sum_{i=0}^{n} \frac{(-1)^i}{i!} \approx 1 - \frac{1}{e} = 0.6321206\ldots,$$

  can be obtained using the inclusion-exclusion rule; try it yourself, then google "matching problem" if you get stuck.

Probabilities can be based on counting outcomes

Example: Toss a coin three times
Define X =Number of Heads in 3 tosses.

	$X(\Omega) = \{0, 1, 2, 3\}$
$\Omega = \{HHH,$	$\longleftarrow$ 3 Heads
$HHT, HTH, THH,$	$\longleftarrow$ 2 Heads
$HTT, THT, TTH,$	$\longleftarrow$ 1 Head
$TTT\}$	$\longleftarrow$ 0 Heads
$p(\{\omega\}) = 1/8$ for each $\omega \in \Omega$	$f(0) = 1/8,\ f(1) = 3/8,$ $f(2) = 3/8,\ f(3) = 1/8$

NB: Even though there are four possible values of X, this fact alone does not imply they all have probability 1/4. **Probability by counting only works when the things being counted are equally likely.**

Pause to summarize

- Probability intuition is very useful, though not infallible
- If sample space is discrete (finite or countably infinite), mathematics is straightforward.
- Any…
  - …random variable is a map from sample space to reals;
  - …event is a subset of the sample space;

  and sometimes conversely.

Outline

There's an intuitive notion of "conditional" probability

- Let X be the number that appears on one die.
- Suppose I peek and tell you that X is even.

- What's the probability that the outcome is in $\{1, 2, 3\}$?
- What's the probability that the outcome is in $\{1, 2, 3, 4\}$?

For events A and B, the probability that A occurs *given that B occurs* is called the conditional probability of A given B and is written $P(A \mid B)$.

Def'n of conditional probability matches intuition

In general, when A and B are events such that $P(B) > 0$, the conditional probability of A given that B has occurred, $P(A \mid B)$, is defined by

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

Specific case: Uniform probabilities

Let $|A| = k$, $|B| = \ell$, $|A \cap B| = j$, $|\Omega| = m$. Given that B has happened, the new probability assignment gives a probability $1/\ell$ to each of the outcomes in B. Out of these ℓ outcomes of B, $|A \cap B| = j$ outcomes also belong to A. Hence

$$P(A \mid B) = j/\ell.$$

Noting that $P(A \cap B) = j/m$ and $P(B) = \ell/m$, it follows that

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

The multiplicative law: Deceptively simple

- When A and B are events such that $P(B) > 0$, the conditional probability of A given that B has occurred, $P(A \mid B)$, is defined by

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

- This definition may be phrased as the *Multiplicative law of conditional probability* (NB: Nothing about independence here!),

$$P(A \cap B) = P(A \mid B)P(B).$$

- The multiplicative law has a generalization to n events:

$$\begin{aligned} P(A_1 \cap A_2 \cap \ldots \cap A_n) &= P(A_n \mid A_1, \ldots, A_{n-1}) \\ &\quad \times P(A_{n-1} \mid A_1, \ldots, A_{n-2}) \\ &\quad \times \ldots \times P(A_2 \mid A_1) \times P(A_1). \end{aligned}$$

The well-known birthday problem uses conditioning

What is the probability that, in a group of n people, there is at least one pair of matching birth dates?

- The answer, given simplifying assumptions, is one minus

$$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \quad (n \text{ terms}). \qquad (1)$$

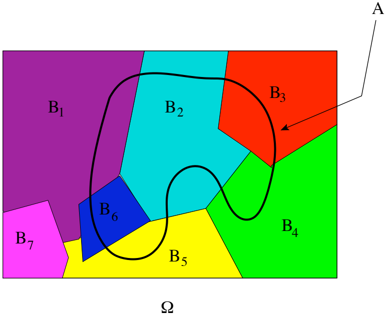
- But why do we multiply these probabilities?
- Answer: If $A_1, A_2, A_3, \ldots$ are chosen cleverly, (1) is simply

$$P(A_1^c) \times P(A_2^c \mid A_1^c) \times P(A_3^c \mid A_1^c, A_2^c) \times \cdots ,$$

which equals $P(A_1^c \cap A_2^c \cap A_3^c \cap \cdots)$.

Partitioning the sample space yields the LOTP

Let $B_1, \ldots, B_k$ be a partition of the sample space Ω (a *partition* is a set of disjoint sets whose union is Ω), and let A be an arbitrary event:



Then

$$P(A) = P(A \cap B_1) + \cdots + P(A \cap B_k).$$

This is called the **Law of Total Probability.** Also, we know that $P(A \cap B_i) = P(A|B_i)P(B_i)$, so we obtain an alternative form of the Law of Total Probability:

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_k)P(B_k).$$

The Law of Total Probability: An Example

Suppose a bag has 6 one-dollar coins, exactly one of which is a trick coin that has both sides HEADS. A coin is picked at random from the bag and this coin is tossed 4 times, and each toss yields HEADS.

Two questions which may be asked here are

- What is the probability of the occurrence of $A = \{\text{all four tosses yield HEADS}\}$?
- Given that A occurred, what is the probability that the coin picked was the trick coin?

The Law of Total Probability: An Example

- What is the probability of the occurrence of $A = \{\text{all four tosses yield HEADS}\}$?

This question may be answered by the Law of Total Probability. Define events

$$\begin{aligned} B &= \text{coin picked was a regular coin,} \\ B^c &= \text{coin picked was a trick coin.} \end{aligned}$$

Then B and B^c together form a partition of Ω. Therefore,

$$\begin{aligned} P(A) &= P(A \mid B)P(B) + P(A \mid B^c)P(B^c) \\ &= \left(\frac{1}{2}\right)^4 \times \frac{5}{6} + 1 \times \frac{1}{6} = \frac{7}{32}. \end{aligned}$$

[Is it clear why $P(A \mid B) = (1/2)^4$?]

The Law of Total Probability: An Example

- Given that A occurred, what is the probability that the coin picked was the trick coin?

For this question, we need to find

$$\begin{aligned} P(B^c \mid A) &= \frac{P(B^c \cap A)}{P(A)} = \frac{P(A \mid B^c)P(B^c)}{P(A)} \\ &= \frac{1 \times \frac{1}{6}}{\frac{7}{32}} = \frac{16}{21}. \end{aligned}$$

Note that this makes sense: We should expect, after four straight heads, that the conditional probability of holding the trick coin, 16/21, is greater than the *prior* probability of 1/6 before we knew anything about the results of the four flips.

Bayes' Theorem: Updating prior prob'ties using data

Suppose we have observed that A occurred.

- Let $B_1, \ldots, B_m$ be all possible scenarios under which A may occur, where $B_1, \ldots, B_m$ is a partition of the sample space.
- To quantify our suspicion that B_i was the cause for the occurrence of A, we would like to obtain $P(B_i \mid A)$.
- Here, we assume that finding $P(A \mid B_i)$ is straightforward for every i. (In statistical terms, a *model* for how A relies on B_i allows us to do this.)
- Furthermore, we assume that we have some prior notion of $P(B_i)$ for every i. (These probabilities are simply referred to collectively as our *prior*.)

Bayes' theorem is the prescription to obtain the quantity $P(B_i \mid A)$. It is the basis of Bayesian Inference. Simply put, our goal in finding $P(B_i \mid A)$ is to determine how our observation A modifies our probabilities of B_i.

Bayes' Theorem: Updating prior prob'ties using data

Straightforward algebra reveals that

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)} = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{m} P(A \mid B_j)P(B_j)}.$$

The above identity is what we call *Bayes' theorem*.

Thomas Bayes (?)

Note the apostrophe *after* the "s". The theorem is named for Thomas Bayes, an 18th-century British mathematician and Presbyterian minister.

The authenticity of the portrait shown here is a matter of some dispute.

Bayes' Theorem: Updating prior prob'ties using data

Straightforward algebra reveals that

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)} = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{m} P(A \mid B_j)P(B_j)}.$$

The above identity is what we call *Bayes' theorem*.

Observing that the denominator above does not depend on i, we may boil down Bayes' theorem to its essence:

$$\begin{aligned} P(B_i \mid A) &\propto P(A \mid B_i) \times P(B_i) \\ &= \text{"the likelihood (i.e., the model)"} \times \text{"the prior"}. \end{aligned}$$

Since we often call the left-hand side of the above equation the *posterior* probability of B_i, Bayes' theorem may be expressed succinctly by stating that the posterior is proportional to the likelihood times the prior.

Bayes' Theorem: Updating prior prob'ties using data

Straightforward algebra reveals that

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)} = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{m} P(A \mid B_j)P(B_j)}.$$

The above identity is what we call *Bayes' theorem*.

There are many controversies and apparent paradoxes associated with conditional probabilities. The root cause is sometimes incomplete specification of the conditions in a particular problem, though there are also some "paradoxes" that exploit people's seemingly inherent inability to modify prior probabilities correctly when faced with new information (particularly when those prior probabilities happen to be uniform).

Try googling "three card problem," "Monty Hall problem," or "Bertrand's box problem" if you're curious.

Independence: Special case of the Multiplicative Law

- Suppose that A and B are events such that
$$P(A \mid B) = P(A).$$
In other words, the knowledge that B has occurred has not altered the probability of A.
- The Multiplicative Law of Probability tells us that in this case,
$$P(A \cap B) = P(A)P(B).$$
- When this latter equation holds, A and B are said to be *independent* events.

Note: The two equations here are not quite equivalent, since only the second is well-defined when B has probability zero. Thus, typically we take the second equation as the mathematical definition of independence.

Independence for more than two events is tricky

- It is tempting but not correct to attempt to define *mutual independence* of three or more events A, B, and C by requiring merely

  $$P(A \cap B \cap C) = P(A)P(B)P(C).$$

  However, this equation does **not** imply that, e.g., A and B are independent.
- A sensible definition of mutual independence should include pairwise independence.
- Thus, we define mutual independence using a sort of recursive definition:

  > *A set of n events is mutually independent if the probability of its intersection equals the product of its probabilities* **and** *if all subsets of this set containing from 2 to $n-1$ elements are also mutually independent.*

Independence for more than two R.V.'s is *not* tricky

- Let X, Y, Z be random variables. Then X, Y, Z are said to be *independent* if

$$\begin{aligned} &P(X \in S_1 \text{ and } Y \in S_2 \text{ and } Z \in S_3) \\ &\qquad = P(X \in S_1)P(Y \in S_2)P(Z \in S_3) \end{aligned}$$

  for **all** possible measurable subsets (S_1, S_2, S_3) of $\mathbb{R}$.
- This notion of independence can be generalized to any finite number of random variables (even two).
- Note the slight abuse of notation:

$$\text{“}P(X \in S_1)\text{” means “}P(\{\omega \in \Omega : X(\omega) \in S_1\})\text{”.}$$

Pause to summarize

- Again: Intuition about conditional probability is very useful.
- The multiplicative law of CP is deceptively powerful.
- Bayes' Theorem basically says

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

- Bayes' Theorem uses data, via the likelihood, to update our prior understanding.

Outline

Expectation of a R.V. is a weighted average

- Let X be a random variable taking values $x_1, x_2 \ldots, x_n$. The expected value μ of X (also called the mean of X), denoted by $E(X)$, is defined by

$$\mu = E(X) = \sum_{i=1}^{n} x_i P(X = x_i).$$

  Note: Sometimes physicists write $\langle X \rangle$ instead of $E(X)$, but we will use the more traditional statistical notation here.

- If $Y = g(X)$ for a real-valued function $g(\cdot)$, then, by the definition above,

$$E(Y) = E[g(X)] = \sum_{i=1}^{n} g(x_i) P(X = x_i).$$

  Generally, we will simply write $E[g(X)]$ without defining an intermediate random variable $Y = g(X)$.

Variance of a R.V.: expected mean squared deviation

- The variance σ^2 of a random variable is defined by

$$\sigma^2 = \mathrm{Var}(X) = E\left[(X - \mu)^2\right].$$

- Using the fact that the expectation operator is linear — i.e.,

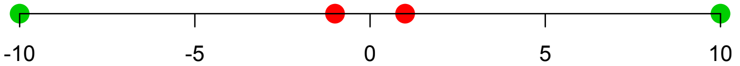
$$E(aX + bY) = aE(X) + bE(Y)$$

  for any random variables X, Y and constants a, b — it is easy to show that

$$E\left[(X - \mu)^2\right] = E(X^2) - \mu^2.$$

- This latter form of $\mathrm{Var}(X)$ is usually easier to use for computational purposes.

Variance measures spread: An example



- Let X be a random variable taking values $+1$ or -1 with probability $1/2$ each.
- Let Y be a random variable taking values $+10$ or -10 with probability $1/2$ each.
- Then both X and Y have the same mean, namely 0, but a simple calculation shows that $\mathrm{Var}(X) = 1$ and $\mathrm{Var}(Y) = 100$.

This simple example illustrates that the variance of a random variable describes in some sense how spread apart the values taken by the random variable are.

Not All Random Variables Have An Expectation

As an example of a random variable with no expectation, suppose that X is defined on some (infinite) sample space Ω so that for all positive integers i,

$$X \text{ takes the value } \begin{cases} 2^i & \text{with probability } 2^{-i-1} \\ -2^i & \text{with probability } 2^{-i-1}. \end{cases}$$

Both the positive part and the negative part of X have infinite expectation in this case, so $E(X)$ would have to be $\infty - \infty$, which is impossible to define.

Counting successes for fixed # trials: Binomial R.V.

- Consider n independent trials where the probability of "success" in each trial is $p \in (0, 1)$; let X denote the total number of successes.
- Then $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \ldots n$.
- X is said to be a *binomial* random variable with parameters n and p, and this is written as $X \sim B(n, p)$.
- One may show that $E(X) = np$ and $Var(X) = np(1-p)$.
- See, for example, A. Mészáros, "On the role of Bernoulli distribution in cosmology," *Astron. Astrophys.*, 328, 1-4 (1997). In this article, there are n uniformly distributed points in a region of volume $V = 1$ unit. Taking X to be the number of points in a fixed region of volume p, X has a binomial distribution. More specifically, $X \sim B(n, p)$.

Counting successes for limitless # trials: Poisson R.V.

- Consider a random variable Y such that for some $\lambda > 0$,

$$P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}$$

  for $y = 0, 1, 2, \ldots$.
- Then Y is said to be a *Poisson* random variable, written $Y \sim \text{Poisson}(\lambda)$.
- Here, one may show that $E(Y) = \lambda$ and $Var(Y) = \lambda$.
- If X has Binomial distribution $B(n, p)$ with large n and small p, then X can be approximated by a Poisson random variable Y with parameter $\lambda = np$, i.e.

$$P(X \leq a) \approx P(Y \leq a)$$

A Poisson random variable in the literature

See, for example, M. L. Fudge, T. D. Maclay, "Poisson validity for orbital debris ..." *Proc. SPIE*, 3116 (1997) 202-209.

The International Space Station is at risk from orbital debris and micrometeorite impact. How can one assess the risk of a micrometeorite impact?

A fundamental assumption underlying risk modeling is that orbital collision problem can be modeled using a Poisson distribution. " ... assumption found to be appropriate based upon the Poisson ... as an approximation for the binomial distribution and ... that is it proper to physically model exposure to the orbital debris flux environment using the binomial distribution. "

Counting failures before first success: Geometric R.V.

- Consider n independent trials where the probability of "success" in each trial is $p \in (0, 1)$
- Unlike the binomial case in which the number of trials is fixed, let X denote the number of failures observed before the first success.
- Then
$$P(X = x) = (1 - p)^x p$$
for $x = 0, 1, \ldots$.
- X is said to be a *geometric* random variable with parameter p.
- Its expectation and variance are $E(X) = \frac{q}{p}$ and $Var(X) = \frac{q}{p^2}$, where $q = 1 - p$.

Failures before rth success: Negative Binomial R.V.

- Same setup as the geometric, but let X be the number of failures before observing r successes.
- $P(X = x) = \binom{r+x-1}{x} (1-p)^x p^r$ for $x = 0, 1, 2, \ldots$.
- X is said to be a *negative binomial* distributon with parameters r and p.
- Its expectation and variance are $E(X) = \frac{rq}{p}$ and $Var(X) = \frac{rq}{p^2}$, where $q = 1 - p$.
- The geometric distribution is a special case of the negative binomial distribution.
- See, for example, Neyman, Scott, and Shane (1953), On the Spatial Distribution of Galaxies: A specific model, *ApJ* **117**: 92–133. In this article, ν is the number of galaxies in a randomly chosen cluster. A basic assumption is that ν follows a negative binomial distribution.

Outline

Beyond Discreteness: Technical issues arise

- Earlier, a *random variable* was a function from Ω to $\mathbb{R}$.
- For discrete Ω, this definition always works.
- But if Ω is uncountably infinite (e.g., if Ω is an interval in $\mathbb{R}$), we must be more careful:

  Definition: *A function* $X : \Omega \to \mathbb{R}$ *is said to be a* random variable *iff for all real numbers a, the set* $\{\omega \in \Omega : X(\omega) \le a\}$ *is an event.*

- Fortunately, we can easily define "event" to be inclusive enough that the set of random variables is closed under all common operations.
- Thus, in practice we can basically ignore the technical details on this slide!

Beyond Discreteness: Sums become integrals

- The function F defined by

$$F(x) = P(X \leq x)$$

  is called the distribution function of X, or sometimes the cumulative distribution function, abbreviated c.d.f.

- If there exists a function f such that

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad \text{for all } x,$$

  then f is called a density of X.

- **Note:** The word “density” in probability is different from the word “density” in physics.

Beyond Discreteness: Sums become integrals

- The function F defined by

$$F(x) = P(X \leq x)$$

  is called the distribution function of X, or sometimes the cumulative distribution function, abbreviated c.d.f.

- If there exists a function f such that

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad \text{for all } x,$$

  then f is called a density of X.

- **Note:** It is typical to use capital "F" for the c.d.f. and lowercase "f" for the density function (recall that we earlier used f for the probability mass function; this creates no ambiguity because a random variable may not have both a density and a mass function).

Beyond Discreteness: Sums become integrals

- The function F defined by

$$F(x) = P(X \leq x)$$

is called the distribution function of X, or sometimes the cumulative distribution function, abbreviated c.d.f.

- If there exists a function f such that

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad \text{for all } x,$$

then f is called a density of X.

- **Note:** Every random variable has a uniquely defined c.d.f. $F(\cdot)$ and $F(x)$ is defined for all real numbers x.
  In fact, $\lim_{x\to-\infty} F(x)$ and $\lim_{x\to\infty} F(x)$ always exist and are always equal to 0 and 1, respectively.

Beyond Discreteness: Sums become integrals

- The function F defined by

$$F(x) = P(X \leq x)$$

  is called the distribution function of X, or sometimes the cumulative distribution function, abbreviated c.d.f.
- If there exists a function f such that

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad \text{for all } x,$$

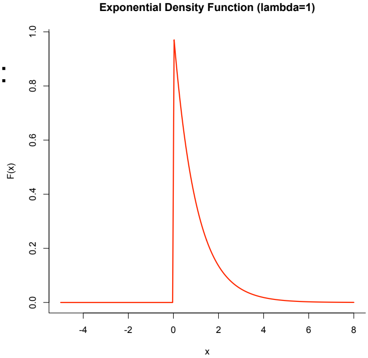
  then f is called a density of X.
- **Note:** Sometimes a random variable X is called "continuous". This does not mean that $X(\omega)$ is a continuous function; rather, it means that $F(x)$ is a continuous function. Thus, it is technically preferable to say "X has a continuous distribution" instead of "X is a continuous random variable."

Continuous version of geometric: Exponential R.V.

- Let $\lambda > 0$ be some positive parameter.
- The *exponential* distribution with mean $1/\lambda$ has desity

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The exponential density for $\lambda = 1$:

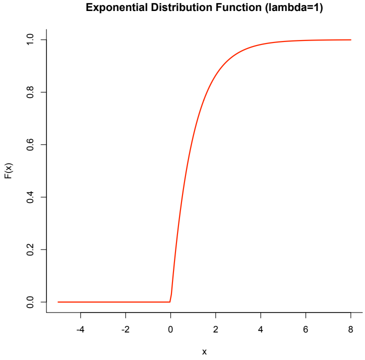


Continuous version of geometric: Exponential R.V.

- Let $\lambda > 0$ be some positive parameter.
- The *exponential* distribution with mean $1/\lambda$ has c.d.f.

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 1 - \exp\{-\lambda x\} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The exponential c.d.f. for $\lambda = 1$:

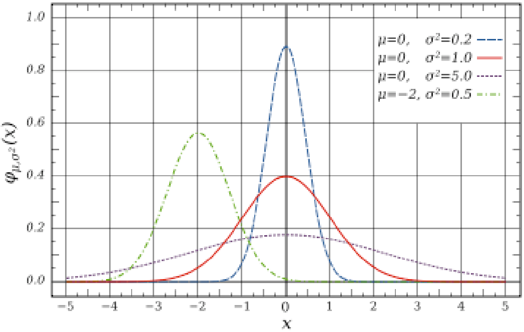


The Normal Distribution

- Let $\mu \in \mathbb{R}$ and $\sigma > 0$ be two parameters.
- The *normal* distribution with mean μ and variance σ^2 has desity

$$f(x) = \varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

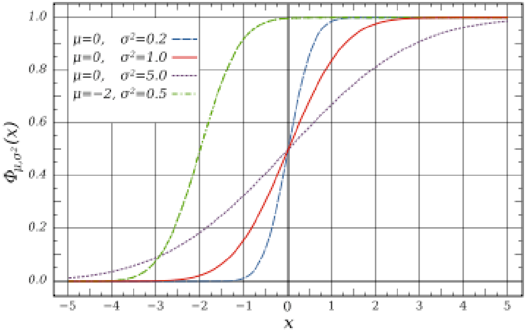
The normal density function for several values of (μ, σ^2):



The Normal Distribution

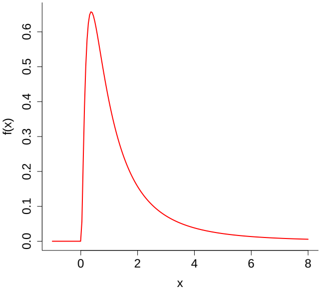
- Let $\mu \in \mathbb{R}$ and $\sigma > 0$ be two parameters.
- The *normal* distribution with mean μ and variance σ^2 has a c.d.f. without a closed form. But when $\mu = 0$ and $\sigma = 1$, the c.d.f. is sometimes denoted $\Phi(x)$. *Remember this notation for later!*

The normal c.d.f. for several values of (μ, σ^2):



The Lognormal Distribution

Standard lognormal density (i.e., $\mu = 0$ and $\sigma = 1$).



- Let $\mu \in \mathbb{R}$ and $\sigma > 0$ be two parameters.
- If $X \sim N(\mu, \sigma^2)$, then $\exp(X)$ has a *lognormal* distribution with parameters μ and σ^2.
- The *lognormal* distribution with parameters μ and σ^2 has density

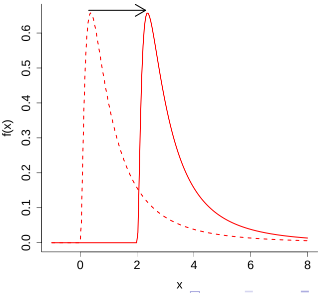
$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right\} \text{ for } x > 0.$$

The Lognormal Distribution

- A common astronomical dataset that can be well-modeled by a shifted lognormal distribution is the set of luminosities of the globular clusters in a galaxy (technically, in this case the size of the shift would be a third parameter).
- With a shift equal to γ, the density becomes

$$f(x) = \frac{1}{(x-\gamma)\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln(x-\gamma)-\mu)^2}{2\sigma^2}\right\} \text{ for } x > \gamma.$$

In the picture, $\gamma = 2$, $\mu = 0$, and $\sigma = 1$.



Expectation: Still a weighted average

- For a random variable X with density f, the expected value of $g(X)$, where g is a real-valued function defined on the range of X, is equal to

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

- Two common examples of this formula are given by the mean of X:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

and the variance of X:

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2.$$

Expectation for two common continuous distributions

- For a random variable X with normal density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

we have $E(X) = \mu$ and $Var(X) = E[(X-\mu)^2] = \sigma^2$.

- For a random variable Y with lognormal density

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right\} \text{ for } x > 0,$$

We have

$$E(X) = e^{\mu+(\sigma^2/2)}$$

$$\mathrm{Var}(X) = E[(X-\mu)^2] = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}.$$

Pause to summarize

- Expectations of RVs preserve summation.
- Variances of *independent* RVs preserve summation.
- There are various parametric families of RV distributions;
  - Sometimes the science suggests a particular family.
  - Sometimes a particular family simply fits the data well.
- Given a parametric family, probability is the study of how RVs behave for given values of the parameters.
- Statistics is probability in reverse: Given observed RVs (data), what can we learn about the parameter values?

Outline

Limit Theorems: Behavior of sums as $n \to \infty$

- Define a random variable X on the sample space for some experiment such as a coin toss.
- When the experiment is conducted many (say, n) times, we are generating a sequence of random variables.
- If the experiment never changes and the results of one experiment do not influence the results of any other, this sequence is called independent and identically distributed (i.i.d.).

Typical notation:

"Let $X_1, \ldots, X_n$ be i.i.d. random variables from some unknown distribution $F(x)$..."

Limit Theorems: Behavior of sums as $n \to \infty$

I win one dollar $\longrightarrow$

$\longleftarrow$ I lose one dollar

- Suppose we gamble on the toss of a coin as follows:
  HEADS means you give me 1 dollar;
  TAILS means you give me -1 dollar, i.e., I give you 1 dollar.
- After n flips, we have $X_1, \ldots, X_n$, where the X_i are i.i.d. and

$$X_i = \begin{cases} +1 & \text{with prob. } 1/2 \\ -1 & \text{with prob. } 1/2. \end{cases}$$

- Then $S_n = X_1 + X_2 + \cdots + X_n$ represents my gain after playing n rounds of this game. We will discuss some of the properties of this S_n random variable.

Limit Theorems: Behavior of sums as $n \to \infty$

I win one dollar $\longrightarrow$

$\longleftarrow$ I lose one dollar

- Recall: $S_n = X_1 + X_2 + \cdots + X_n$ represents my gain after playing n rounds of this game.
- Here are some possible events and their probabilities. **The proportion of games won is the same in each case.**

OBSERVATION	PROBABILITY
$S_{10} \leq -2$ i.e. I lost at least 6 out of 10	0.38 moderate
$S_{100} \leq -20$ i.e. I lost at least 60 out of 100	0.03 unlikely
$S_{1000} \leq -200$ i.e. I lost at least 600 out of 1000	1.36×10^{-10} impossible

Law of Large Numbers: Famous, oft misunderstood!

- Suppose $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with $E(X_1) = \mu < \infty$. Then

$$\overline{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}$$

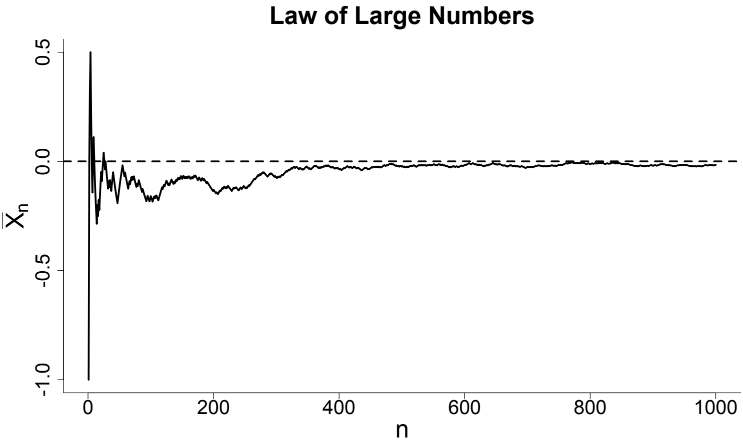
converges to $\mu = E(X_1)$ in the following sense: For any fixed $\epsilon > 0$,

$$P(\mid \overline{X}_n - \mu \mid > \epsilon) \longrightarrow 0 \text{ as } n \to \infty.$$

- In words: The sample mean $\overline{X}_n$ converges to the population mean μ.
- It is very important to understand the distinction between the sample mean, which is a random variable and depends on the data, and the true (population) mean, which is a constant.

Law of Large Numbers: Famous, oft misunderstood!

In our example in which S_n is the sum of i.i.d. ± 1 variables, here is a plot of n vs. $\overline{X}_n = S_n/n$ for a simulation:



Central Limit Theorem: Fundamental Thm of Prob'ty?

- Let $\Phi(x)$ denote the c.d.f. of a standard normal (mean 0, variance 1) distribution.
- Based on 1000 flips, S_{1000} has mean 0 and variance 1000.
- Thus, $S_{1000}/\sqrt{1000}$ has mean 0 and variance 1
- Consider the following table:

Event	Probability	Exact Normal
$S_{1000}/\sqrt{1000} \leq 0$	0.513	$\Phi(0) = 0.500$
$S_{1000}/\sqrt{1000} \leq 1$	0.852	$\Phi(1) = 0.841$
$S_{1000}/\sqrt{1000} \leq 1.64$	0.947	$\Phi(1.64) = 0.950$
$S_{1000}/\sqrt{1000} \leq 1.96$	0.973	$\Phi(1.96) = 0.975$

- It seems as though $S_{1000}/\sqrt{1000}$ behaves a bit like a standard normal random variable.

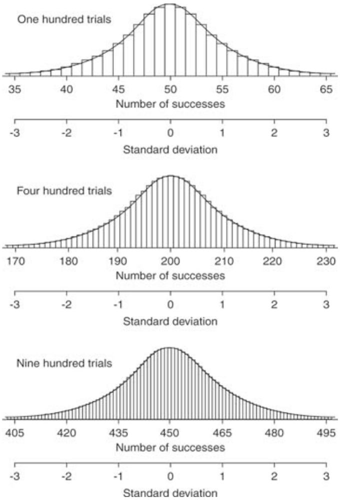
Central Limit Theorem: Fundamental Thm of Prob'ty?

- Suppose $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables such that $E(X_1^2) < \infty$.
- Let $\mu = E(X_1)$ and $\sigma^2 = E[(X_1 - \mu)^2]$. In our coin-flipping game, $\mu = 0$ and $\sigma^2 = 1$.
- Let
$$\overline{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}.$$
- Remember: μ is the population mean and $\overline{X}_n$ is the sample mean.
- Then for any real x,
$$P\left\{\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \leq x\right\} \longrightarrow \Phi(x) \text{ as } n \to \infty.$$

This (**amazing!**) fact is called the *Central Limit Theorem*.

Central Limit Theorem: Fundamental Thm of Prob'ty?

The CLT is illustrated by the following figure, which gives idealized histograms based on the coin-flipping game:



Still more structure: Law of the Iterated Logarithm

With probability 1,

$$\limsup_n \sqrt{n}\,\overline{X}_n = -\liminf_n \sqrt{n}\,\overline{X}_n = \infty.$$

But if we divide by $\sqrt{2\log\log n}$, we get

$$\limsup_n \frac{\sqrt{n}}{\sqrt{2\log\log n}}\overline{X}_n = -\liminf_n \frac{\sqrt{n}}{\sqrt{2\log\log n}}\overline{X}_n = 1 \quad \text{wp } 1.$$