The University of Texas at Austin
School of Information

Stats: Probability

Mick McQuaid

2024-02-05

The University of Texas at Austin
School of Information

Week FOUR

Probability is hard!

Example of probability

Rolling a pair of dice

- You know a general rule
- You want to know a single outcome (or a small set of outcomes)

Example of stats

A/B testing a website

- You know a single outcome (or a small set of outcomes)
- You want to know a general rule (or an appropriate model)

In a way, stats and probability are inverses of each other

Approaches to probability

Bayesian interpretation

There are two competing schools of thought about what probability is. The Bayesian approach is that probability is *quantified belief or reasonable expectation of the outcomes of events based on a state of knowledge*. This approach is recently taught in graduate schools. It requires a lot of math background and is not widely studied in undergraduate courses. We will not extensively study this approach, but I want you to know that it exists and is rising in academic popularity.

Frequentist interpretation

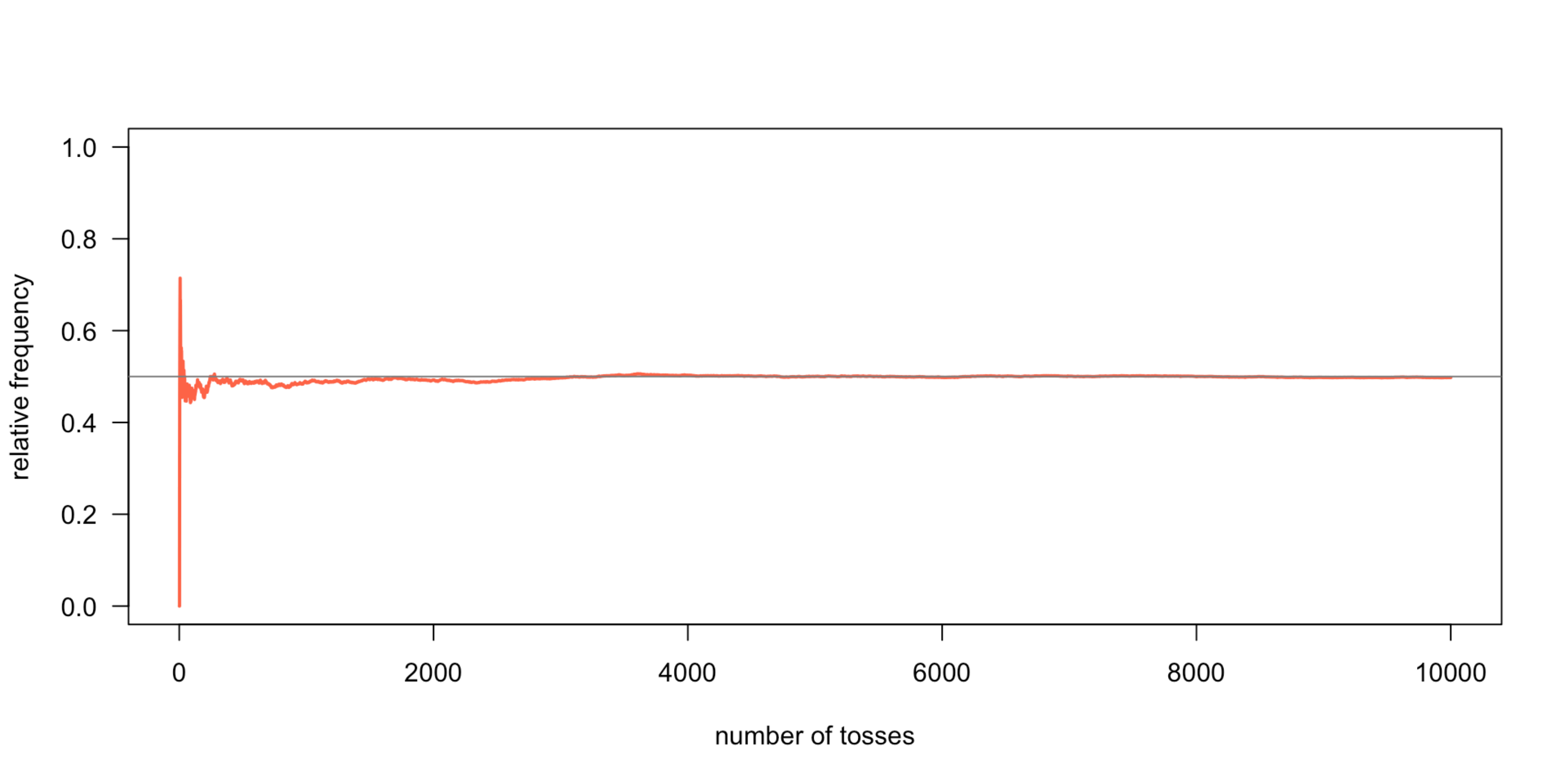
Our textbook takes a *frequentist* approach to probability, one of the two main approaches to probability and the one usually taught in undergraduate courses in the USA. This approach models probability of an *outcome* as the number of times the outcome would occur if we observed the *random process* that produced it an infinite number of times. For example, if we flip a fair coin an infinite number of times, it comes up heads half the time, so the probability of heads is 0.5.

Law of large numbers

This law claims that, as more outcomes are observed, the proportion of outcomes converges to the probability of the outcome. For example, if we flip a fair coin a hundred times, the probability of heads coming up half the time is greater than if we only flip it ten times.

Example of law of large numbers

```
flips
heads tails
 4978  5022
```



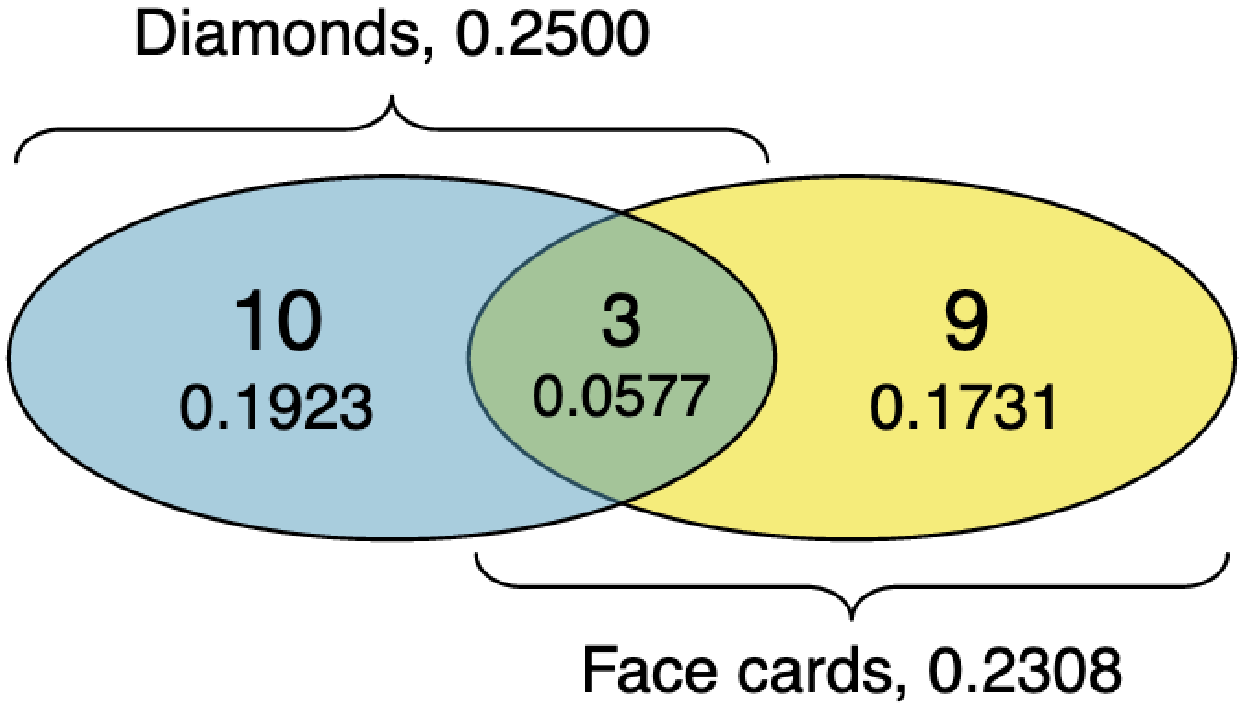
Outcomes—disjoint and otherwise

Disjoint outcomes

These are outcomes that can not both happen. For example, in the fair coin flipping case, the outcome cannot be both heads and tails. But the sum of all the disjoint probabilities is always 1.

Probabilities when outcomes are not disjoint

The textbook uses playing cards to illustrate concepts like *cards that are neither diamonds nor face cards*. You have to familiarize yourself with playing cards to understand these examples. The textbook uses the following Venn diagram to illustrate the above example.



There are also 30 cards that are neither diamonds nor face cards

Figure 3.4: A Venn diagram for diamonds and face cards.

General addition rule

The textbook gives a general rule for multiple outcomes, whether they are disjoint or not.

If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

where $P(A \text{ and } B)$ is the probability that both events occur.

Counting Permutations

Permutations can be thought of as lineups. For instance, suppose you have five people to put in a line. There are five people to choose from to be first in line, then four people remain to be second in line, and so on. You can count this up as $5 \times 4 \times 3 \times 2 \times 1 = 5!$ or five factorial. This holds true for as many objects as you wish to line up.

Counting Combinations

Combinations can be thought of as committees. There is no order as in a lineup. You’re either a member or you’re not. Suppose you want to form a committee of five people from among twenty people. It doesn’t matter what order they come in so you can’t use the factorial method to count them. Another method is shown in Ash (1993). The main result is that, to choose a committee of 5 from among 20 people, use

$$\binom{20}{5} = \frac{20!}{5!(20-5)!}$$

This is read as *twenty choose 5*.

The binomial coefficient

The definition of the binomial coefficient is

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

Rules for combinations

$$\binom{n}{r} = \binom{n}{n-r}$$

$$\binom{n}{1} = n$$

$$\binom{n}{n} = \binom{n}{0} = 1$$

This last result is because $0! = 1$ by definition.

Example of a combination

Ash (1993) gives the examples of finding and not finding the Queen of Spades (Q_s) in a poker hand. You can think of a poker hand as a committee of 5 cards drawn from 52, so the total number of poker hands is given by $\binom{52}{5}$. Finding hands containing the Q_s amounts to choosing a committee of size four (the remainder of the hand, from among 51 cards (the remainder of the deck. So there are $\binom{51}{4}$ such hands.

$$P(Q_s) = \frac{\binom{51}{4}}{\binom{52}{5}} = \frac{5}{52}$$

Simplifying stacked fractions

You need to do this when solving a combination by hand.

$$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

Probability distributions

Probability distribution definition

A probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

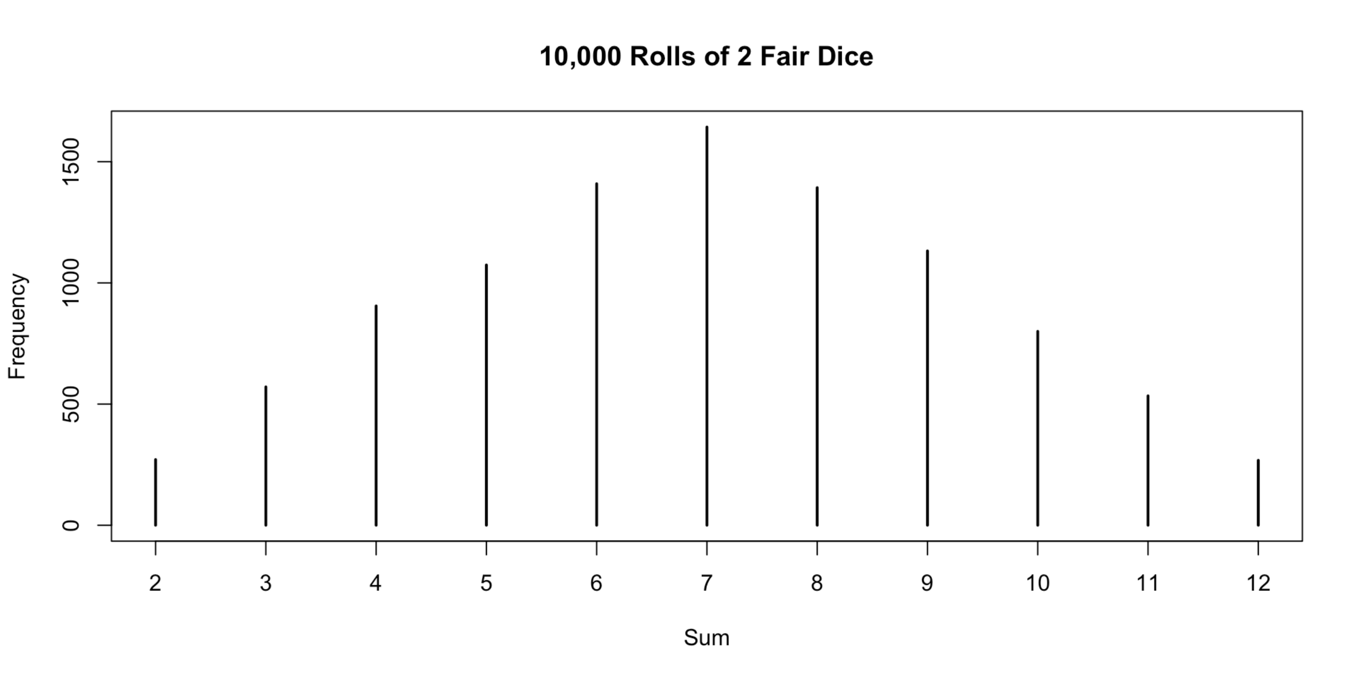
1. The outcomes listed must be disjoint.

2. Each probability must be between 0 and 1.

3. The probabilities must total 1.

Probability distribution for two fair dice

Francis DiTraglia shows the following example of plotting the probability distribution for rolling two fair dice on his website.



Why is 7 the most likely outcome?

Terms from set theory

Set theory is a mathematical discipline that uses tools like Venn diagrams to describes sets of objects. We can think of outcomes from random processes as objects, too, with the following terms.

- sample space: the set of all possible outcomes
- event: a particular outcome
- complement of an event: outcomes in the sample space outside a given event or events

Complementary events

The complement of event A is denoted A^c, and A^c represents all outcomes not in A. A and A^c are mathematically related:

$$P(A) + P(A^c) = 1, \text{ i.e., } P(A) = 1 - P(A^c)$$

Independence

Just as variables and observations can be independent, random processes can be independent, too. Two processes are *independent* if knowing the outcome of one provides no useful information about the outcome of the other. For instance, flipping a coin and rolling a die are two independent processes—knowing the coin was heads does not help determine the outcome of a die roll. On the other hand, stock prices usually move up or down together, so they are not independent.

Multiplication rule for independent processes

If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$P(A \text{ and } B) = P(A) \times P(B)$$

Similarly, if there are k events $A_1, \dots, A_k$ from k independent processes, then the probability they all occur is

$$P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

Conditional probability

Conditional probability example

This is where probability gets interesting. Some things depend on other things! The textbook uses a contingency table of the `photos_classify` data frame, which you can download from OpenIntro Stats, to explore this concept.

```
              truth
mach_learn     fashion  not  Sum
  pred_fashion     197   22  219
  pred_not         112 1491 1603
  Sum              309 1513 1822
```

We can use the entries in the contingency table to make statements about probability.

- probability that a fashion photo is correctly classified by ML (machine learning): 197/309
- probability that a given photo is about fashion when predicted by ML to be not: 112/1603

Marginal probability

Marginal probability is the probability in the margins of table (right column and bottom row), e.g., ML predicts fashion photo at all: 219/1822.

Joint probability

Joint probability is the probabillity of two (or more) things being true, e.g., ML predicts fashion and truth is fashion: 197/1822. A joint probability would be any of the four interior cells divided by the lower right cell.

Conditional probability

Conditional probability is the probability of some outcome given the condition of another outcome, such as the probability that ML predicts fashion when the photo is truly about fashion: 197/219.

Conditional probability's importance

Conditional probability is very important. It's useful to know the general formula for conditional probability. The conditional probability of outcome A given condition B is computed as the following:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

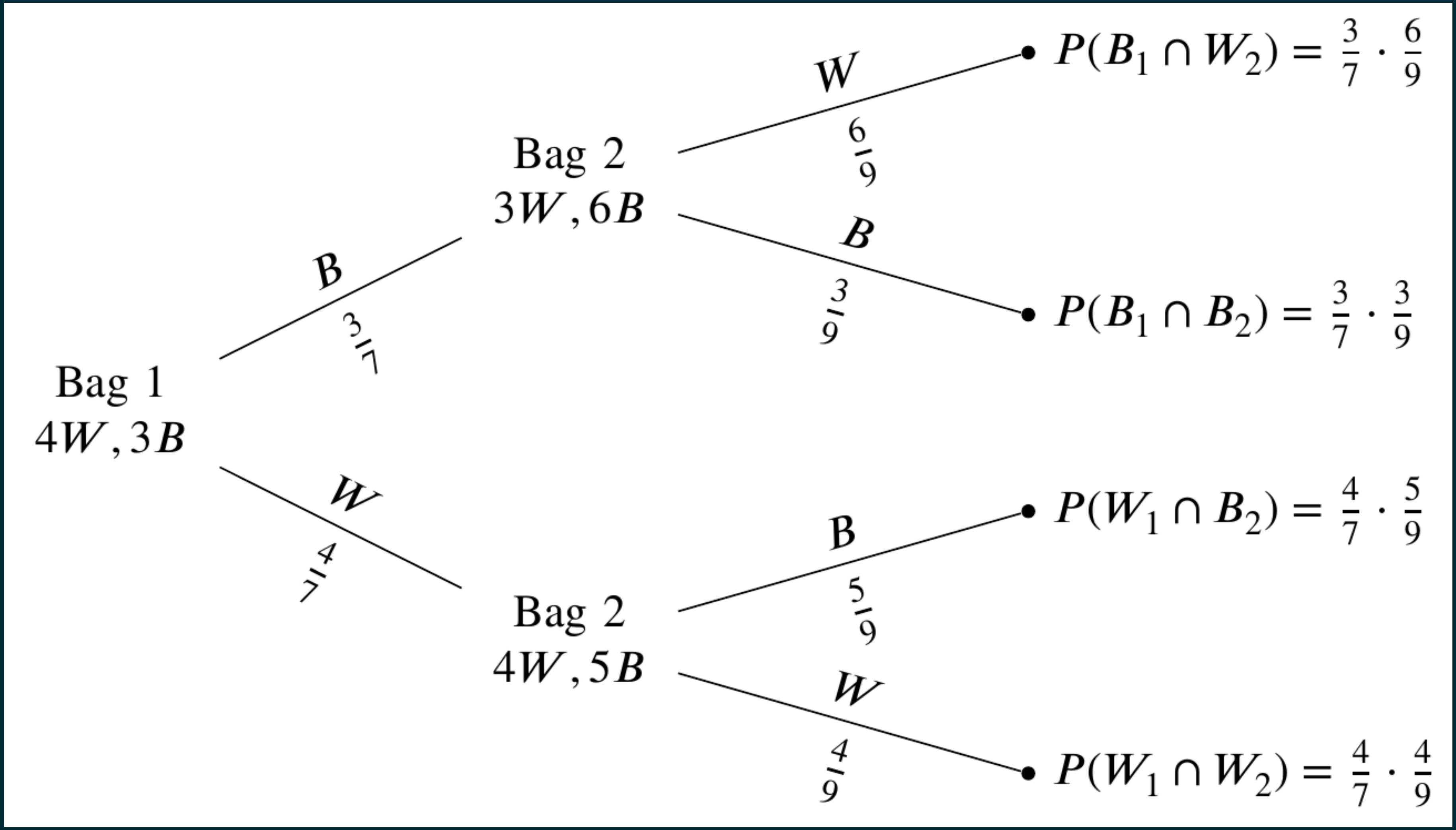
General multiplication rule

We already saw a specific multiplication rule for independent events. But there is a more general rule, applicable whether independence is true or not. If A and B represent two outcomes or events, then

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

It is useful to think of A as the outcome of interest and B as the condition.

Tree diagrams


Bag 1
$4W, 3B$
B
$\frac{3}{7}$
W
$\frac{4}{7}$
Bag 2
$3W, 6B$
Bag 2
$4W, 5B$
W
$\frac{6}{9}$
B
$\frac{3}{9}$
B
$\frac{5}{9}$
W
$\frac{4}{9}$
$P(B_1 \cap W_2) = \frac{3}{7} \cdot \frac{6}{9}$
$P(B_1 \cap B_2) = \frac{3}{7} \cdot \frac{3}{9}$
$P(W_1 \cap B_2) = \frac{4}{7} \cdot \frac{5}{9}$
$P(W_1 \cap W_2) = \frac{4}{7} \cdot \frac{4}{9}$

Textbook tree diagram

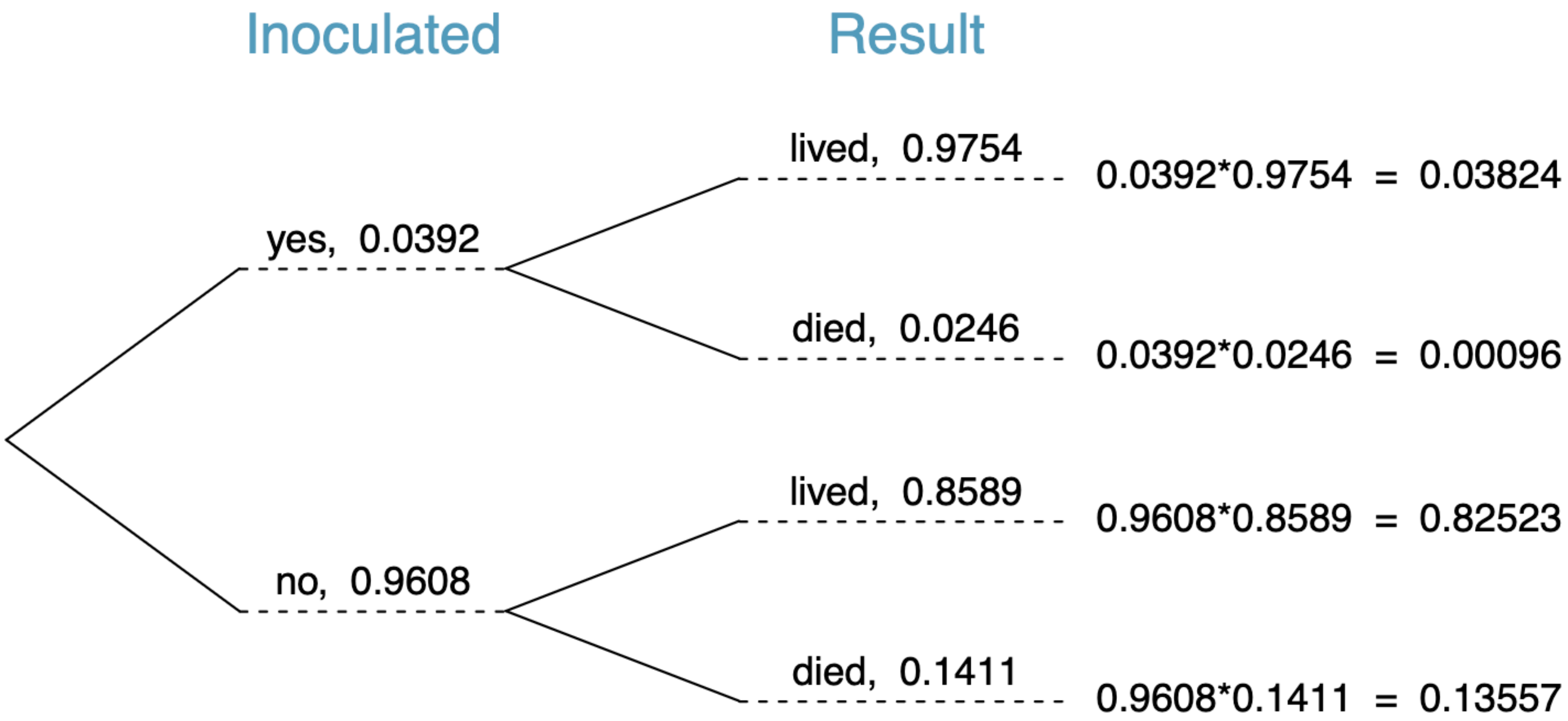


Figure 3.17: A tree diagram of the smallpox data set.

Bayes’ Theorem

Bayes’ Theorem Definition

It’s the centerpiece of Bayesian statistics so it’s a bit like a fish out of water in a frequentist course. It states, in words, that the posterior probability of an outcome A given an outcome B is the likelihood of the outcome B times the prior probability of the outcome A divided by the evidence of outcome B. More succinctly,

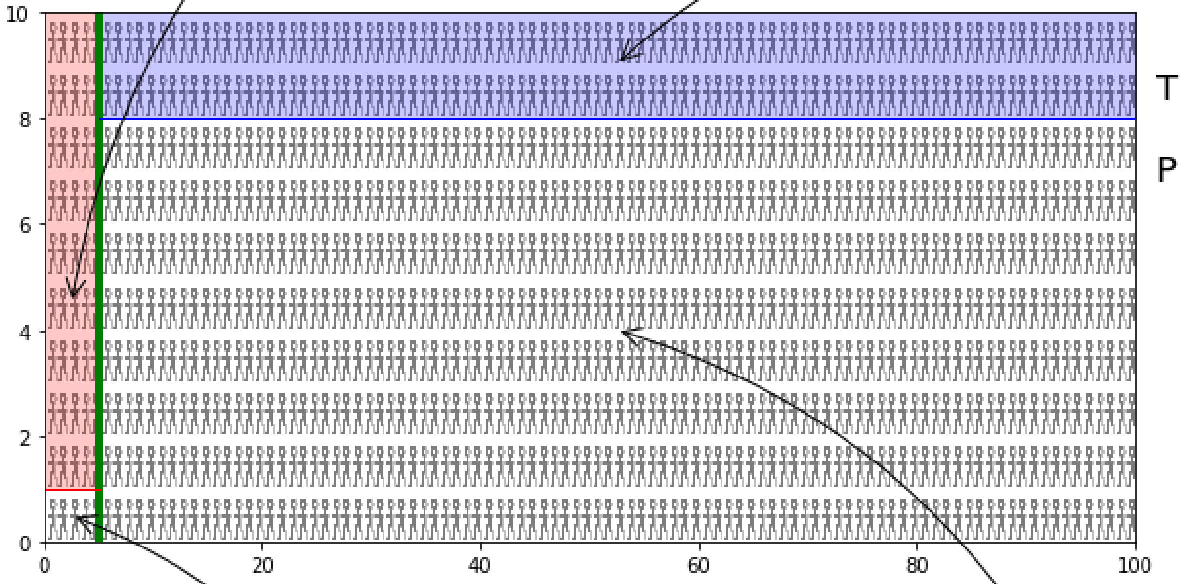
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes rule intuition

The intuition is that we usually know *something* and shouldn’t go into problem solving with no assumptions. An example of drug user testing given by Wikipedia is shown in the following frame graphically, where the test is ninety percent sensitive to a recipient being a drug user. The test is also eighty percent specific, meaning that it can detect that a non-user is a non-user eighty percent of the time.

Users and test Positive:
45 out of 50 Users

Non-users and test Positive:
190 out of 950 Non-Users



The total number of people tested is 1,000.

Prob(User|Positive) is 19%

= 45 out of 235 people testing Positive

= The ratio of the pink area to the combined pink and blue areas

Non-Users and test Negative:
760 out of 950 Non-Users

Users but test Negative:
5 out of 50 Users

Bayes rule conclusion

What does this tell us about drug testing? Even if someone tests positive, the probability that they are a drug user is only 19%! This assumes prior knowledge that five percent of the general population are users of the drug.

Drug Testing in Bayesian Terms

$$\frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.2 \times 0.95}$$

By the way, this formulation uses the law of total probability in the denominator, expanding $P(B)$ into

$$P(B|A)P(A) + P(B|\neg A)P(\neg A)$$

where $\neg A$ is the complement of A.

Bayes Theorem Rationale

This is very important to know about because, for example, someone might *only* tell you that a test is ninety percent sensitive and eighty percent specific and leave out the Bayes rule result that says that, given that only five percent of the general population are drug users, there’s a nineteen percent chance that testing positive indicates that you are a drug user. Again, Bayes’ rule is very important when you have some prior knowledge. In this case, the prior knowledge is a study that says that five percent of the general population are users of this drug.

Small populations

Sampling from a small population

Recall that the population includes every object and is usually not practical to measure. For examples, every fish, every person, every thunderstorm are too many to represent. So we take a sample. A typical rule of thumb (or heuristic) is that, if we can sample more than ten percent of the population, we regard it as a small population.

Small population example

The textbook gives an example of being called on by the professor. The chance that you are called on in a class of 15 is 1/15. If the professor calls on three different people in succession, the chance that you are called on increases to 1/5:

$$\begin{aligned} P(\neg 3 \text{ in a row}) &= \\ &= P(\text{not picked first, second, third}) \\ &= \frac{14}{15} \times \frac{13}{14} \times \frac{12}{13} \\ &= \frac{12}{15} \end{aligned}$$

and the complement of 12/15 is 1/5.

Random variables

Random variable definition

A process with a random numerical outcome is called a random variable. It's kind of like a stochastic function in that there is an input (the process) and output (the outcome number).

A random variable is usually represented as a capital, italicized Latin letter, e.g., X, Y, Z. Specific outcomes are usually represented as a lowercase, italicized Latin letter with a subscript to denote which outcome, such as x_1 , x_2 , x_3 . The probability that a random variable X has a specific outcome x_1 is represented as $P(X = x_1)$.

Expectation

The expectation of X is the expected value of X, represented as $E(X)$. The expected value is typically the average, but not always. If there are k possible outcomes, then

$$E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

Σ

The Greek letter $\sum$ (Sigma) denotes a sum of a series of numbers, indexed in this case by i. You can read it in English as the sum, going from 1 to k, of the expressions to the right of the Sigma sign. In other words,

$$x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_k P(X = x_k)$$

It's the average as we usually understand it if each outcome is equally probable, in which case it's the sum of the outcomes divided by the number of outcomes.

μ instead of $E(X)$

Writers often substitute $E(X) = \mu$ which can be confusing because Greek letters are usually used as parameters, while Latin letters are usually used as specific realizations of those parameters.

The above assumes there are k specific outcomes. We call this case a *discrete random variable*. It's also possible to do math with a *continuous random variable*. In other words, $k = \infty$. However, that requires calculus so the textbook skips it for now.

Variability in random variables

Recall that random variables are a kind of mapping between a process and specific numerical outcomes. Those specific outcomes differ. For example, the revenue of a store varies day by day, and we can say something about that variability.

Variance and standard deviation

We usually express it by two related concepts: variance (denoted by a lowercase sigma squared or σ^2), and its square root, standard deviation (denoted by a lowercase sigma or σ). You might wonder why we don't just choose one of these symbols. Most statisticians just use standard deviation, but to prove that it is an unbiased estimator of variance, we need to square it for mathematical reasons that are beyond the scope of this course.

We prefer to use standard deviation

Standard deviation is expressed in the same units as the subject under consideration, whereas variance is expressed in squared units. So when I said in the study guide that the standard deviation of completed exercises was about 2.69, I meant that in terms of number of exercises, meaning that most of you were within 2.69 completed exercises either way of each other. In other words, if you had three or nine completed exercises, you were a kind of outlier.

Linear combinations of random variables

We can put random variables together. For example, your GPA is calculated from a set of grades that may differ. If you play fantasy sports, your score comes from many different players. A recommender system for music listening may make calculations based on many different songs you listened to.

Expressing linear combinations

A linear combination of two random variables can be expressed as $aX + bY$, where a and b are fixed constants, for example, the number of credits applied to each class in your GPA portfolio. If you take four four credit classes, your GPA might be expressed as the linear combination

$$4E(X_1) + 4E(X_2) + 4E(X_3) + 4E(X_4)$$

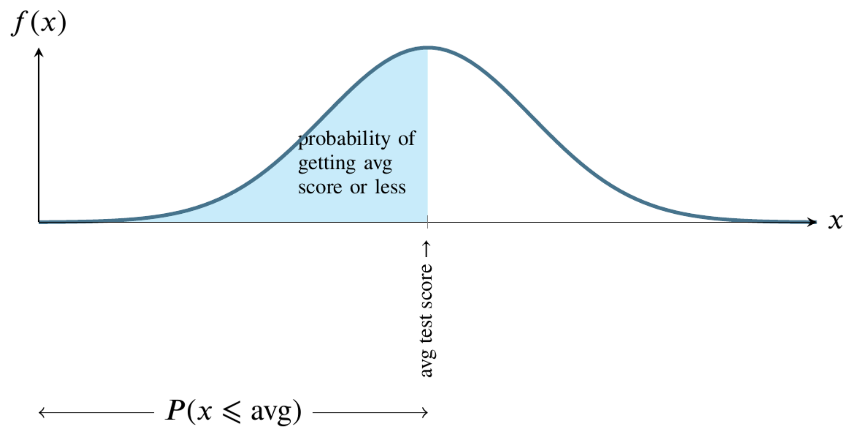
where $E(X_i)$ is a function of your grade, such as your grade mapped to a number and divided by the number of credits you’re taking.

Continuous distributions

So far, we've considered discrete distributions, where k takes on one of a finite set of values. What about the continuous case? For example, temperature can theoretically take on an infinite number of values, even though we only have the tools for discrete measurements.

Normal distribution

The most famous continuous distribution is the normal distribution. This picture illustrates the normal distribution. The mound-shaped curve represents the probability density function and the area between the curve and the horizontal line represents the value of the cumulative distribution function.


$f(x)$
probability of getting avg score or less
x
avg test score →
$P(x \leqslant \text{avg})$

Normal distribution example (1 of 2)

Consider a normally distributed nationwide test.

The total shaded area between the curve and the straight horizontal line can be thought of as one hundred percent of that area. In the world of probability, we measure that area as 1. The curve is symmetrical, so measure all the area to the left of the highest point on the curve as 0.5. That is half, or fifty percent, of the total area between the curve and the horizontal line at the bottom. Instead of saying *area between the curve and the horizontal line at the bottom*, people usually say *the area under the curve*.

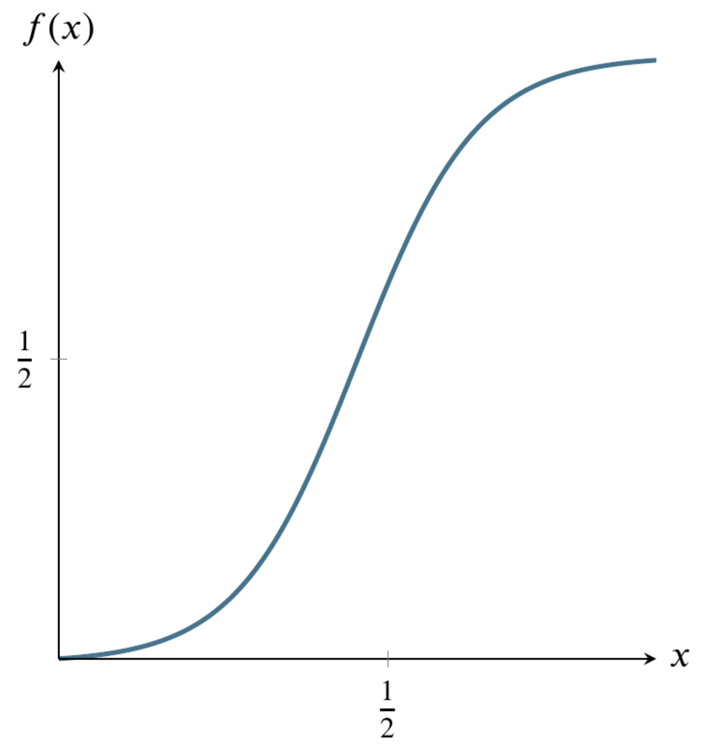
Normal distribution example (2 of 2)

For any value along the *x*-axis, the *y*-value on the curve represents the value of the probability density function.

The area bounded by the vertical line between the *x*-axis and the corresponding *y*-value on the curve, though, is what we are usually interested in because that area represents probability.

The Cumulative Distribution Function (CDF)

Here is a graph of the *size* of that area. It’s called the cumulative distribution function (cdf).



Explanation of the CDF graph

The above graph can be read as having an input and output that correspond to the previous graph of the probability density function. As we move from right to left on the x-axis, the area that would be to the left of a given point on the probability density function is the y-value on this graph. For example, if we go half way across the x-axis of the probability density function, the area to its left is one half of the total area, so the y-value on the cumulative distribution function graph is one half.

Logistic or Sigmoid curves

The shape of the cumulative distribution function is called a sigmoid curve (also sometimes called a logistic curve). You can see how it gets the s shape by looking again at the probability density function graph above. As you move from left to right on that graph, the area under the curve increases very slowly, then more rapidly, then slowly again. The places where the area grows more rapidly and then more slowly on the probability density function curve correspond to the s-shaped bends on the cumulative distribution curve.

Reading the sigmoid curve left to right

At the left side of the cumulative distribution curve, the *y*-value is zero meaning zero probability. When we reach the right side of the cumulative distribution curve, the *y*-value is 1 or 100 percent of the probability.

The nationwide test again

Let's get back to the example of a nationwide test. If we say that students nationwide took an test that had a mean score of 75 and that the score was normally distributed, we're saying that the value on the x-axis in the center of the curve is 75. Moreover, we're saying that the area to the left of 75 is one half of the total area. We're saying that the probability of a score less than 75 is 0.5 or fifty percent. We're saying that half the students got a score below 75 and half got a score above 75.

Frequentism

That is called the frequentist interpretation of probability. In general, that interpretation says that a probability of 0.5 is properly measured by saying that, if we could repeat the event enough times, we would find the event happening half of those times.

Furthermore, the frequentist interpretation of the normal distribution is that, if we could collect enough data, such as administering the above test to thousands of students, we would see that the graph of the frequency of their scores would look more and more like the bell curve in the picture, where x is a test score and y is the number of students receiving that score.

Measuring areas under the bell curve

Suppose we have the same test and the same distribution but that the mean score is 60. Then 60 is in the middle and half the students are on each side. That is easy to measure. But what if, in either case, we would like to know the probability associated with scores that are not at that convenient midpoint?

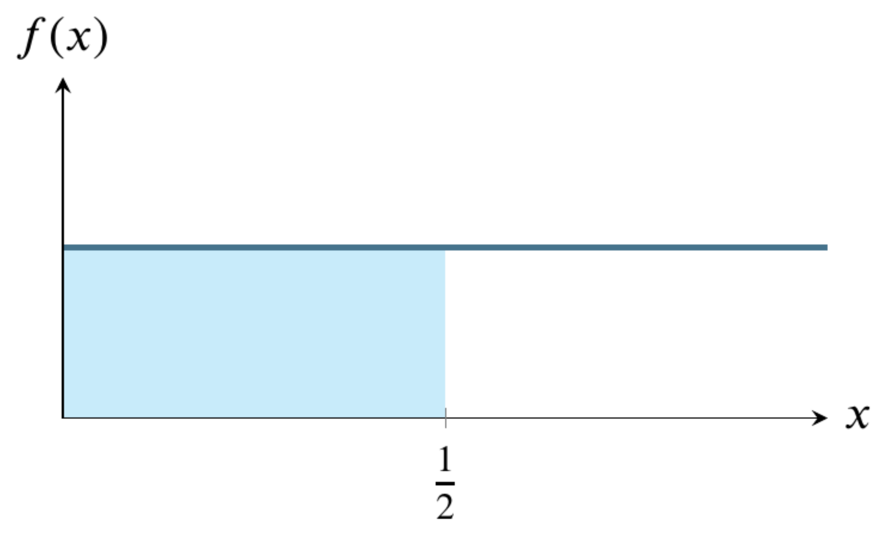
It's hard to measure any other area under the normal curve except for x-values in the middle of the curve, corresponding to one half of the area. Why is this?

The uniform distribution

To see why it’s hard to measure the area corresponding to any value except the middle value, let’s first consider a different probability distribution, the uniform distribution. Suppose I have a machine that can generate any number between 0 and 1 at random. Further, suppose that any such number is just as likely as any other such number.

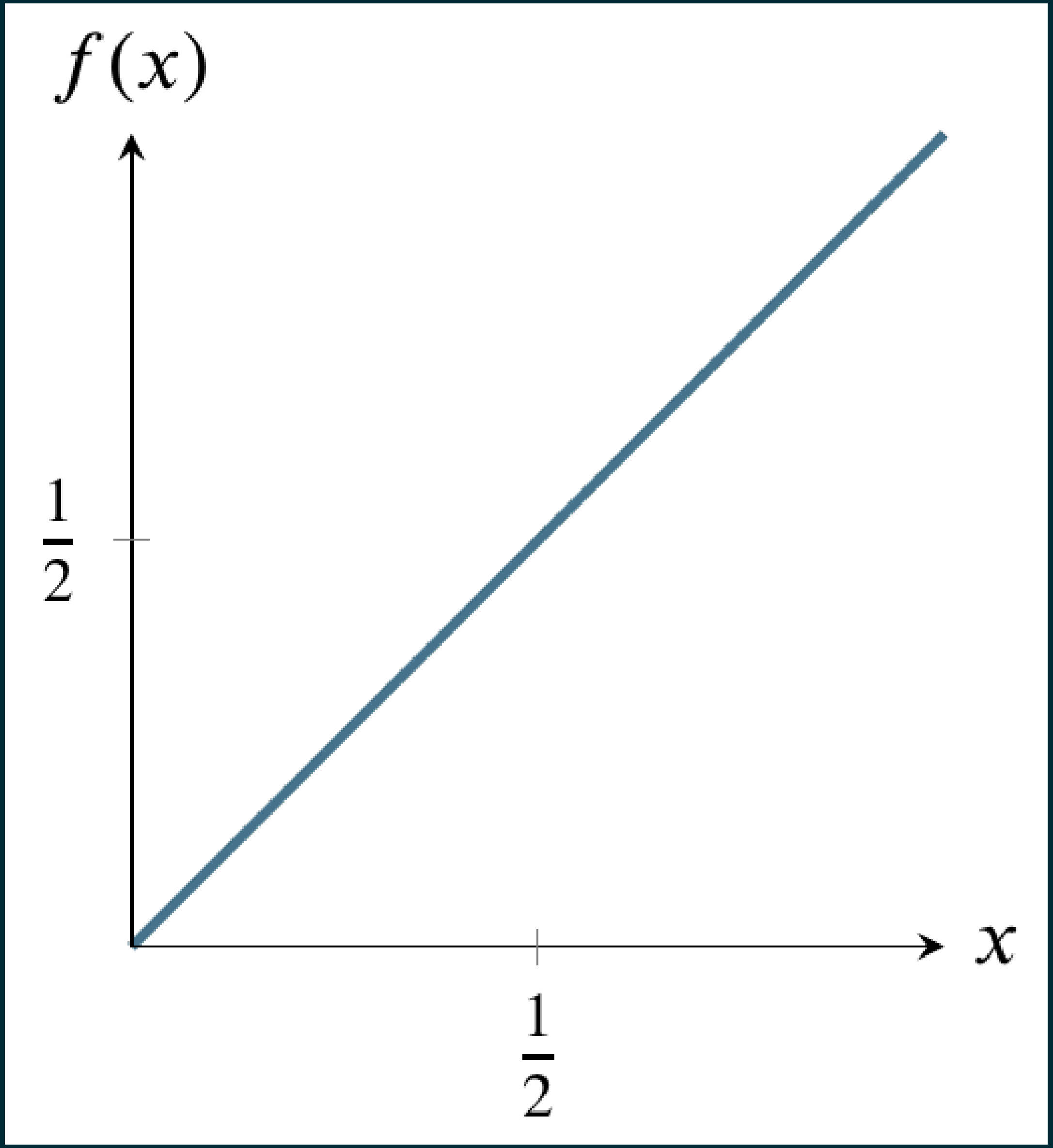
Graphing the uniform distribution

Here's a graph of the the uniform distribution of numbers generated by the machine. The horizontal line is the probability density function and the shaded area is the cumulative distribution function from 0 to 1/2. In other words, the probability of the machine generating numbers from 0 to 1/2 is 1/2.



Areas under the uniform curve

It's very easy to calculate any probability for this distribution, in contrast to the normal distribution. The reason it is easy is that you can just use the formula for the area of a rectangle, where area is base times side. The probability of being in the entire rectangle is $1 \times 1 = 1$, and the probability of being in the part from $x = 0$ to $x = 1/4$ is just $1 \times (1/4) = 1/4$. The cumulative distribution function of the uniform distribution is simpler than that of the normal distribution because area is being added at the same rate as we move from left to right on the above graph. Therefore it is just a straight diagonal line from (0,1) on the left to (1,1) on the right.


$f(x)$
$\frac{1}{2}$
x
$\frac{1}{2}$

Reading it is the same as reading the cumulative distribution function for the normal distribution. For any value on the *x*-axis, say, 1/2, go up to the diagonal line and over to the value on the *y*-axis. In this case, that value is 1/2. That is the area under the horizontal line in the probability density function graph from 0 to 1/2 (the shaded area). For a rectangle, calculating area is trivial.

Calculating a curved region

Calculating the area of a curved region like the normal distribution can be more difficult. If you’ve studied any calculus, you know that there are techniques for calculating the area under a curve. These techniques are called integration techniques.

Integration techniques

In the case of the normal distribution the formula for the height of the curve at any point on the x-axis is

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

and the area is the integral of that quantity from $-\infty$ to x, which can be rewritten as

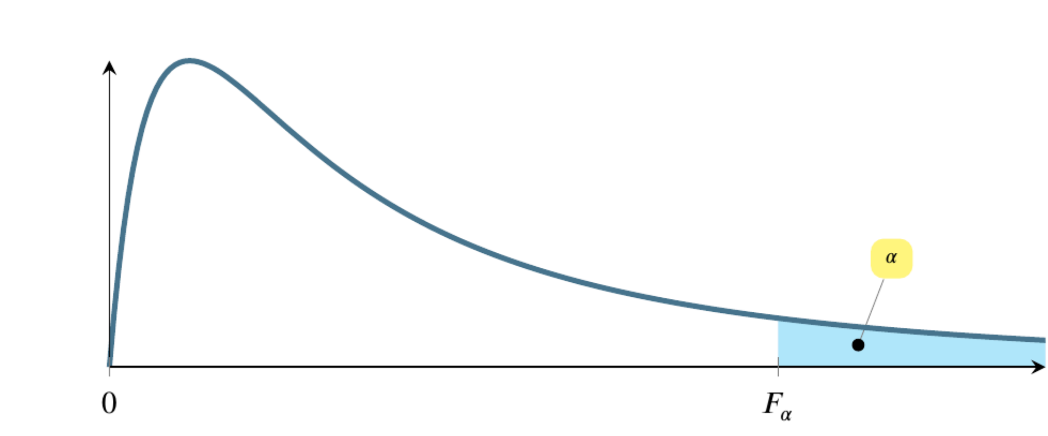
$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt = (1/2)\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

Numerical approximation

The integral on the left is difficult to evaluate so people use numerical approximation techniques to find the expression on the right in the above equation. Those techniques are so time-consuming that, rather than recompute them every time they are needed, a very few people used to write the results into a table and publish it and most people working with probability would just consult the tables. Only in the past few decades have calculators become available that can do the tedious approximations. Hence, most statistics books, written by people who were educated decades ago, still teach you how to use such tables. There is some debate as to whether there is educational value in using the tables vs using calculators or smartphone apps or web-based tables or apps

The F distribution

Another distribution that we'll enounter later is called the F distribution. We'll calculate an F-statistic when we build linear regression models, but I just want you to know the general shape of it for now. The region marked α corresponds inversely to the magnitude of the F statistic. In other words, a larger F statistic means a smaller α.



END

References

Ash, Carol. 1993. *The Probability Tutoring Book*. New York, NY: IEEE Press.

Colophon

This slideshow was produced using `quarto`

Fonts are *Roboto Condensed Bold*, *JetBrains Mono Nerd Font*, and *STIX2*