Introduction to probability and distributions, Part 1

XDASI Fall 2021

9/20/2021

Contents

Reading assignment

- **Whitlock & Schluter, Chapter 5: Probability**
- **Optional**: ***Aho, Foundational and Applied Statistics for Biologists with R***
  - **Chapter 2**
  - **Chapter 3.1-3.2**

Classical Probability

At its core, statistics is about probabilities and probabilistic inference. So we need to build some common vocabulary to talk about probability.

Terminology

Let's start with some of the basics:

⇒ ***Question: What is a variable?***

Answer

- Something whose value can change.

For example, the formula for instantaneous velocity of a falling body is:

$$f(t) = gt$$

- where $g = 9.8m/s^2$ is acceleration due to gravity (a constant), and
- $t = time(sec)$ (a variable)

Deterministic vs. probabilistic models

⇒ ***Q: What is a deterministic model?***

Answer

- Something that when given the same inputs, will always produce the same output.
- Such models are useful conceptually, but almost never characterize real-world phenomena.

⇒ ***Q: What is a probabilistic model?***

Answer

- A model that incorporates **random variables**.
- Such models need not produce **exact** outputs, but the **most probable** outcome.

For example, a probabilistic model for the instantaneous velocity of a falling body is:

$$f(t) = gt + \varepsilon$$

- where ε represents measurement error, and
- the expected value of ε is zero: $\bar{\varepsilon} = 0$.
- Values of ε that are close to zero will occur with the highest frequency.
- This value represents the **maximum likelihood**.

Random variables and random trials

When we do experiments with an unknown outcome, we need a common framework for thinking about the process and outcomes. Let's review some basic concepts:

⇒ ***Q: What is a random variable?***

Answer

- Something whose value cannot be known preceding a measurement (a.k.a. **trial**).

⇒ ***Q: What is a random trial?***

Answer

- A process with two or more possible outcomes that are not predictable.

⇒ ***Q: What is an event?***

Answer

- An observed outcome of a random trial.

⇒ ***Q: What is a probability, in terms of random trials?***

Answer

- The proportion of random trials with a particular outcome.
- The probability always between zero and one: $0 \le P \le 1$

In mathematical notation, we can write:

$$P[A] = \frac{N(A)}{N}$$

Which is to say, "the ***probability*** of event A" equals the ***proportion*** of events with outcome A relative to all possible outcomes.

Set theory and notation

- Set
- Element
- Subset
- Sample space
- Event
- Empty set
- Probability of an event
- Proportion
- Sample space
- Null set

Relationships between outcomes

- Disjoint sets (mutual exclusion)
  - Intersect
  - Union (addition rule)
  - Example: Blood type
- Nondisjoint sets
  - Intersect
  - Union (general addition rule)
  - Example: Blood type
- Independence
  - Multiplication rule
  - Addition
  - Can mutually exclusive outcomes be independent?
  - How common is independence?

Probability distribution functions

A ***probability distribution*** describes the probabilities of all possible outcomes of a random variable.

Quantitative random variables

Discrete vs. **continuous** distributions describe situations in which there are a ***limited*** vs. an ***infinite*** number of outcomes.

- A **probability density function (PDF)** is a mathematical expression that defines a probability distribution.
- A *discrete* probability distribution is technically called a **probability mass function (PMF)**, but we will commonly refer to these as PDFs.

Discrete distributions are generally illustrated using **histograms**, whereas continuous PDFs are illustrated using line graphs:

Some terminology

To write mathematical formulas for distributions, we need some notation to describe random variables, outcomes, and probability distributions:

- A random variable, X
- A (continuous) random variable outcome, x
- Discrete variable outcomes are called *mass points* x_i
- A PDF, $f(x)$
- The output generated by a PDF, a *density*
- A cumulative distribution function (CDF), $F(x)$

Density

Discrete and continuous probability distributions both generate a quantity called the *density*. The *density function*, given as $f(x)$, gives the **height** of a PDF for any outcome x.

Both types (discrete and continuous) will be valid iff ("if and only if"):

1. $f(x) \geq 0, \forall x \in \mathbb{R}$
2. $\sum_x f(x) = 1$ (discrete PDF) or $\int_{-\infty}^{\infty} f(x)dx = 1$ (continuous PDF)

In other words, the density function must be zero or positive for all possible outcomes, and the total probability of all possible outcomes must equal one. Thus, **the total area under any probability density always equals one**. PDFs can be represented by histograms or, for continuous densities, as curves.

Distribution functions

In R, there are four families of commands relating to distributions that you should become familiar with. For a *normal* distribution, these are:

- **rnorm**: generates *random samples* from the normal distribution
- **dnorm**: gives the *density function (PDF)*
- **pnorm**: gives the *cumulative distribution function (CDF)*
- **qnorm**: gives the inverse CDF, a.k.a. the *quantile function*

You will learn how to use all of these.

The PDF

The PDF answers the question, for a particular probability distribution, "What is the probability of observing a value **exactly equal to** x as an outcome of the random variable X?" It is the *relative frequency* of a particular outcome, given all possible outcomes.

For a **discrete PDF**, all possible values of x are *discrete*, so the density at any point x_i is equivalent to a *probability*. Thus we can write:

$$f(x) = P(X = x), \quad x \in X = \{x_1, x_2, ...\}.$$

Note that this is *not true* for *continuous* functions, since for a continuous variable the probability at any discrete value of x is zero (it is necessary to integrate across some interval to get a finite area under the curve).

For continuous distributions, we will see the density may sometimes exceed one across a range of values. Nevertheless, **the total area under any PDF is always equal to one**.

The cumulative distribution function (CDF)

The CDF answers the question, “What is the probability of observing a value **less than or equal to** x as an outcome?” This is called a *lower-tail probability.*

It can also be used to answer the question, “What is the probability of observing a value **greater than** x?” This is the *upper-tail probability* and is obtained by subtracting the value of the CDF at x from 1.

The CDF for a random variable X is denoted $F(x)$ and gives the **lower-tail probability** $P(X \leq x)$ for the corresponding PDF. This probability is given by the **total area** underneath a density, for all outcomes **up to and including** the value x.

For a **discrete** random variable:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

For a **continuous** random variable:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Note that we find the total probability for a continuous random variable using integration, which gives us the area under a curve. To be proper, we called the continuous variable of integration in the above formula t, since we are using it to find x. This is just a formality. The point is that the total probability of getting a value at least as big as x is the area under the curve from minus infinity up to x. Closed forms of some continuous PDFs allow solutions to be found without the need for integration.

Fortunately, many distribution functions are already built into R, so we don’t usually have to worry about integration! (R also provides the capability to perform integration directly, if you want to check that the built-in functions are giving you the correct answer.)

Discrete distributions

The Bernoulli distribution

The Bernoulli distribution describes the **probability of success for a single trial of a binary random variable**. If we encode the outcomes as either 1 (“success”) or 0 (“failure”), then we can write a formula for this as:

$$f(x) = P(X = x) = \pi^x(1-\pi)^{1-x}, \quad x \in \{0, 1\}$$

where π represents the probability of “success”, and ranges from zero to one: $0 \leq \pi \leq 1$.

The formal way of writing this function seems kind of complicated, but it’s really pretty simple. Since x can only take on values of 0 or 1, and there is only one trial, $f(x)$ can take on only one of two values:

$$f(1) = P(X = 1) = \pi \quad or \quad f(0) = P(X = 0) = 1 - \pi.$$

The CDF is:

$$F(X) = \begin{cases} 1 - \pi & x = 0 \\ 1 & x = 1 \end{cases}$$

What does this mean in practice? For a fair coin toss, the chance of getting heads or tails is the same: $P(X=1) = P(X=0) = 0.5$.

For another example, let's say you are trying to fuse a GFP tag onto the end of a CDS in *C. elegans* using CRISPR, and the efficiency is around 20% (this may be unrealistically high, depending on how well CRISPR is currently working in your lab, but let's go with it anyway). This means that if you pick only one worm, the chance of recovering a line WITH the GFP tag is 20%, and the chance of that worm NOT having the GFP tag is 80%. So, $P(X=1) = 0.2$, and $P(X=0) = 0.8$.

The Binomial distribution

What if you are not just interested in a single Bernoulli trial, but you want to know how many worms you will have to pick to get at least three independent GFP lines, or to have an 80% chance of getting a GFP transformant? Being able to figure out the answer to this kind of question can help you plan your experiments better.

This is what the binomial distribution is for! It one of the most fundamental distributions in probability theory. The binomial distribution gives the probability of a particular number of "successes" x, given n i.i.d. ("independent and identically distributed") Bernoulli trials with a fixed probability π of success for each trial.

The binomial distribution is a function of two variables, n and π, and we denote it as $X \sim BIN(n, \pi)$. The tilde means that the random variable X "follows" the binomial distribution. The binomial probability mass function is:

$$f(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad x \in \{0, 1, ..., n\}$$

The above equation simply says that, in order to find the total probability for a particular outcome, we need to multiply the probability of the outcome we observe by the total number of ways that this can happen.

Let's break this equation down into its component parts:

1. The term $\binom{n}{x}$ is called the *binomial coefficient* and is spoken as "n choose x". It is **number of possible combinations** of x successes (and $n-x$ failures) out of n Bernoulli trials.
2. The rest of the equation is just the **probability for one of these combinations**: $\pi^x(1-\pi)^{n-x}$.
    - Since each trial is independent, we just follow the ***Product Rule*** to find the probability that any number of trials had a particular outcome.
    - Since each trial has π probability of success, the probability of x successes is π^x. For 2 successes, this would be π^2, and so on.
    - Similarly, if there are x successes, then there are $n-x$ failures, each with probability $1-\pi$, so the probability of $n-x$ failures is $(1-\pi)^{n-x}$.

The mean and variance of a Binomial distribution are: $\mu = E(X) = np$ and $\sigma^2 = V(X) = np(1-p)$. The full CDF is:

$$F(X) = \sum_{x=0}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x} = 1$$

Illustration Let's continue the CRISPR example above. You probably need to pick more than one worm to find your GFP strain! Let's say you pick 3 worms. What's the probability that two out of the three will be transformants?

First, let's ask how likely it is that the first two worms you pick will be transformants, and the third will not? Well, that works out to $(0.2) * (0.2) * (0.8) = 0.032$.

But, you're not done yet! You need to consider ***how many ways*** there are to get 2 out of 3 transformants. Let's work this out using Set Theory:

- There is only one way to get zero transformants in three tries: $S = \{000\}$.

- There are three ways to get one GFP strain and two non-GFP strains: $S = \{100, 010, 001\}$.
- Similarly, there are 3 ways to get 2 GFP strains and 1 non-GFP strain: $S = \{110, 101, 011\}$.
- Finally, there is only one way to get 3 transformants in three tries: $S = \{111\}$.

This is what the $\binom{n}{k}$ part of the equation is for! Now we are ready to solve the problem: $P(k = 2) = \binom{3}{2}(0.2)^2(0.8) = 3 * 0.032 = 0.096$. If you pick only three worms, you'll have about a 1 in 10 chance of finding exactly two transformants.

You're probably more interested in the chance of finding ***at least*** 2 transformants (it's always good to have more than one independent CRISPR line!). To do this, we will need to use the CDF. Specifically, we are interested in the ***upper-tail*** probability that we will find *more than one* transformant. To do this, we add up the *lower-tail probabilities* for zero or one transformants, and subtract the sum from 1:

$$P(k = 0) = \binom{3}{0}(0.8)^3 = 0.512$$

$$P(k = 1) = \binom{3}{1}(0.2)(0.8)^2 = 3 * 0.128 = 0.384$$

So, $P(X > 1) = 1 - (0.512 + 0.384) = 0.104$. This is slightly better, but not much!

You can compute this using the R function for the PDF. Fortunately, it gives the same result!

```r
# cumulative upper-tail probability of getting MORE THAN one transformant: P(X > 1)
1-pbinom(1,3,0.2)                # 1 minus the lower-tail probability
```

```
## [1] 0.104
```

```r
pbinom(1,3,0.2,lower.tail=FALSE) # this is equivalent
```

```
## [1] 0.104
```

How many worms would you have to pick to guarantee an 80% chance of getting at least two transformants? It would be pretty tedious to calculate this out by hand, especially as the number of trials increases.

```r
# probability of getting at least 2 transformants for different numbers of worms picked
pbinom(1,3,0.2,lower.tail=FALSE)
```

```
## [1] 0.104
```

```r
pbinom(1,5,0.2,lower.tail=FALSE)
```

```
## [1] 0.26272
```

```r
pbinom(1,8,0.2,lower.tail=FALSE)
```

```
## [1] 0.4966835
```

```r
pbinom(1,11,0.2,lower.tail=FALSE)
```

```
## [1] 0.6778775
```

```r
pbinom(1,14,0.2,lower.tail=FALSE) # you should check at least 14 if you want 2 or more!
```

```
## [1] 0.8020879
```

The Binomial Theorem One interesting feature of the Binomial distribution is that it is **symmetric**. This means that the number of ways you can get exactly 2 out of 3 "successes" is the same as the number of ways you can get exactly 1 out of 3 successes. If in the example above $P(success)$ were 0.5 instead of 0.2, then you'd have the same chance of finding exactly one or exactly two transformants, since the chance of success or failure would be the same.

Pascal's Triangle The term "n choose k" has a special name, the **Binomial coefficient**. For any n, it is possible to work out the number of possible ways of achieving any outcome using Pascal's Triangle:

$$\begin{array}{rcl}
\binom{0}{0} & = & 1 \\
\binom{1}{0}\binom{1}{1} & = & 1 \quad 1 \\
\binom{2}{0}\binom{2}{1}\binom{2}{2} & = & 1 \quad 2 \quad 1 \\
\binom{3}{0}\binom{3}{1}\binom{3}{2}\binom{3}{3} & = & 1 \quad 3 \quad 3 \quad 1
\end{array}$$

The top line represents the possible combinations for $n = 0$, the second line for $n = 1$, etc. Notice that the sum of two components at a higher level of Pascal's triangle equals the component of the next lower level that is situated directly beneath and between them.

It is easy to see that the Bernoulli distribution is a special case of the Binomial distribution with $n = 1$. There's only one way to get one success or one failure in one trial!

Binomial Expansion If we consider two Bernoulli trials (where the probability of a success for each trial is π), we can use the product rule to work out the probability of two successes (call these A), two failures (call these B), or one success and one failure:

$$P(A \cap A) = P(A)P(A) = \pi^2$$
$$P(A \cap B) = 2P(A)P(B) = 2\pi(1 - \pi)$$
$$P(B \cap B) = P(B)P(B) = (1 - \pi)^2$$

There is only one way to get two successes or two failures ($S = \{00\}$ or $S = \{11\}$), and two ways to get one of each ($S = \{10, 01\}$). This pattern should look familiar to you; it's the same as for the binomial expansion $(x + y)^n$, where $n = 2$:

$$(x + y)^2 = x^2 + 2xy + y^2$$

Binomial Theorem The expansion above is a special case of the **Binomial Theorem**. This can be extended to any arbitrary number of trials using the product rule for independent events. The binomial coefficients can be found using the "n choose k" shortcut, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Instead of writing out Pascal's Triangle to get the number of possible combinations for each outcome, we just use the general formula for arbitrary n and k:

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

The second expression is equivalent to the first because the Binomial distribution is symmetric.

More generally, we can say that $(x + y)^n$ can be expressed as a sum of terms of the form ax^by^c, where $b + c = n$ and each a is a ***binomial coefficient*** that is a specific positive integer that depends on n and b (or, equivalently, n and c).

In the limit, a binomial distribution looks very much like a normal, or Gaussian distribution. We will convince ourselves of this in a future class exercise.

Examples

Bernoulli distribution

Example: the proportion of green M&Ms in a bag of M&Ms was found to be 0.17. What is the probability that the next M&M you pick from your bag will not be green?

$$P(X = 0) = f(0) = (0.17)^0(1 - 0.17)^1 = 0.83$$

Binomial distribution (example 1)

During the industrial revolution in England, London become covered in a lot of black soot from burning coal. Five years beforehand, the proportion of white moths that could be found in London was 87%.[1] However black moths gained a survival advantage as the air became more polluted.

If you were to sample moths in London 25 years into the industrial revolution, and you found that 35 out of 50 moths were white, how likely would that be if the population had remained the same?

You can compute the probability of finding ***exactly*** 35/50: (Answer: 0.088%)

$$P(X = 35) = \binom{50}{35} 0.87^{35}(1 - .87)^{50-35}$$

Fortunately, there is a function that performs this calculation for us in R:

```
dbinom(35,size=50,prob=0.87) # this is the density function (PDF)
```

```
## [1] 0.000880373
```

Really what you are probably more interested in is finding out how likely it is that you only found ***no more than*** 35 in total. To do this, you'd have to add up all the probabilities of getting 0, 1, 2, ... 35 white moths. This is given by the binomial CDF:

$$P(X \leq 35) = \sum_{x_i \leq 35} f(x_i) = \sum_{x_i \leq 35} \binom{50}{x_i} 0.87^{x_i}(1 - 0.87)^{50-x_i}$$

Wow, that looks nasty! You could use a loop to compute this in R:

```
cdf <- function(x,n,p) {
  result = numeric()
  for (i in 0:x) {
    f = n-i
    q = 1-p
    result[i] = choose(n,i)*(p^i)*(q^f)
  }
  return(sum(result))
}
cdf(35,50,0.87)
```

```
## [1] 0.001285362
```

... but the R function for the binomial CDF is so much simpler! It is:

```
pbinom(35,50,0.87) # cumulative probability for x between 0 and 35
```

```
## [1] 0.001285362
```

[1] Disclaimer: I made this up, but based on a true story (see https://en.wikipedia.org/wiki/Peppered_moth_evolution).

It's hard to tell from just one sample what the true proportion really is . . . we will talk a lot more about this soon! For now, we can estimate how likely it would be that you would find, say, somewhere between 30-40 white moths, vs. 40-50, if the population had not changed:

```
sum(dbinom(30:40,50,0.87))  # very unlikely!
```

```
## [1] 0.1074296
```

```
sum(dbinom(40:50,50,0.87))  # much more likely
```

```
## [1] 0.946505
```

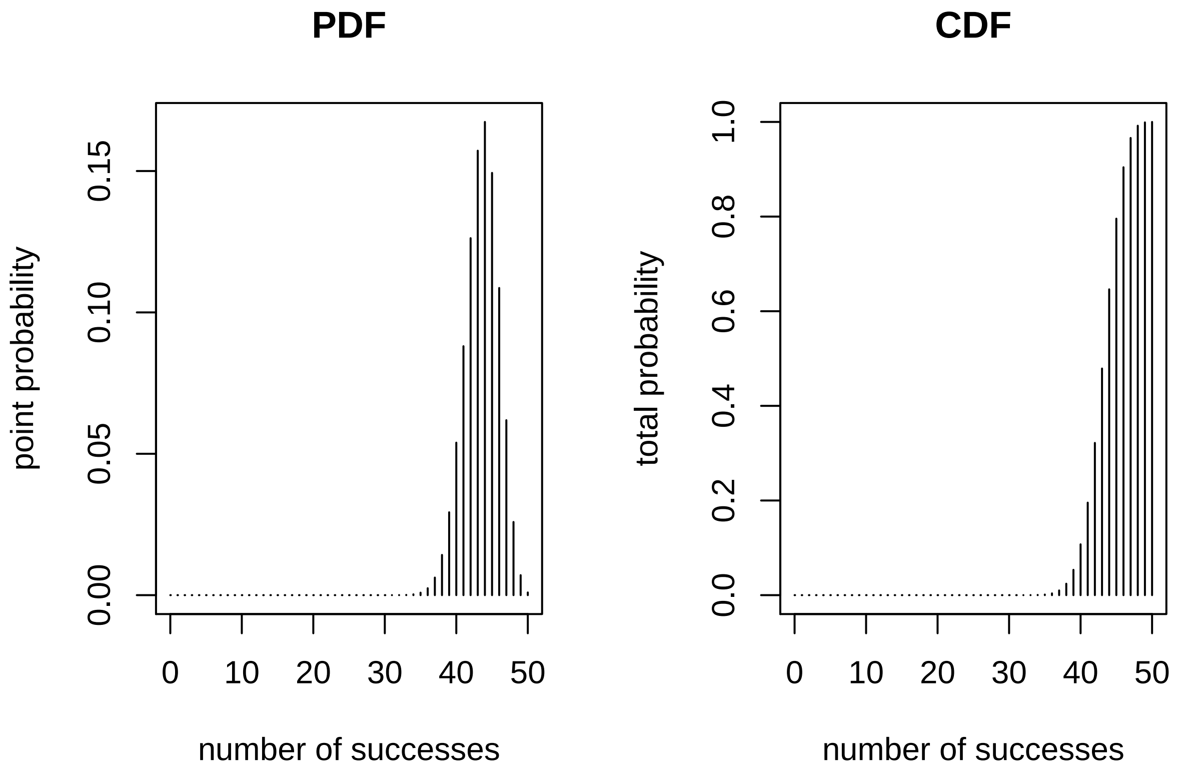
To find the probability of getting 40 or more white moths if the population had stayed the same, i.e. $P(X \geq 40)$, you could also use the CDF. You'd have to add `lower.tail = FALSE` and use 39 instead of 40, or else you would be asking for $P(X > 40)$:

```
pbinom(39,50,0.87,lower.tail=FALSE)
```

```
## [1] 0.946505
```

We can visualize the density and the CDF by plotting them, which makes it a lot easier to understand the probabilities we computed above:

```
par(mfrow=c(1,2))
x = 0:50
plot(x,dbinom(x,size=50,prob=0.87),type="h",main="PDF",
     xlab="number of successes",ylab="point probability")
plot(x,pbinom(x,size=50,prob=0.87),type="h",main="CDF",
     xlab="number of successes",ylab="total probability")
```



Binomial distribution (example 2)

Consider a genome in which the four bases A, C, G, T are present in equal proportions (this is actually rather uncommon, but let's go with it for now).

If you were to pick a sequence of 10nt completely at random, what is the chance that exactly 3 bases will be an A (or any other homopolymer)?

```
dbinom(3,10,.25)
```

```
## [1] 0.2502823
```

What is the chance that you will find less than 5 Gs (i.e. 4 or fewer)?

```
pbinom(4,10,.25)
```

```
## [1] 0.9218731
```

What is the chance that you will find between 2 and 4 Ts?

$$P(2 \leq X \leq 4) = \sum_{k=2}^{4} \binom{10}{k} (0.25)^k (0.75)^{10-k} = \sum_{k=0}^{4} \binom{10}{k} (0.25)^k (0.75)^{10-k} - \sum_{k=0}^{1} \binom{10}{k} (0.25)^k (0.75)^{10-k}$$

```
# these are equivalent: a) sum up the probabilities for all values
sum(dbinom(2:4,10,.25))
```

```
## [1] 0.6778479
```

```
# b) get the cumulative mass up to and including 4, then
#    subtract the cumulative up to 1 (to get 2 inclusive)
pbinom(4,10,.25) - pbinom(1,10,.25)
```

```
## [1] 0.6778479
```

The quantile function `qbinom(p, size, prob)` returns the smallest value of q such that $Pr(X \leq q) \geq p$. The quantile is defined as the smallest value x such that $F(x) \geq p$. For example, $F(x) \geq 0.75$ means that 75% of the distribution is less than x. Here, you are likely to find 3 or fewer of each kind of base in 75% of 10-mers picked at random from the genome:

```
qbinom(.75,10,.25)
```

```
## [1] 3
```

What is the maximum number of any single base you would expect to find in 99% of random 10-mers?

```
qbinom(.99,10,.25)
```

```
## [1] 6
```

⇒ ***Question: ?***

Answer

- Answer.