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: $0 \leq P \leq 1$

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.

Usually, discrete distributions are illustrated using histograms, and continuous PDFs are illustrated using line graphs, for example:

```
# ========================================================================== #
# binomial

## A short function to compute number of ways to get
## a total face value for 2 fair dice, ranging from 2 to 12
ncomb.2dice = function (value) {
```

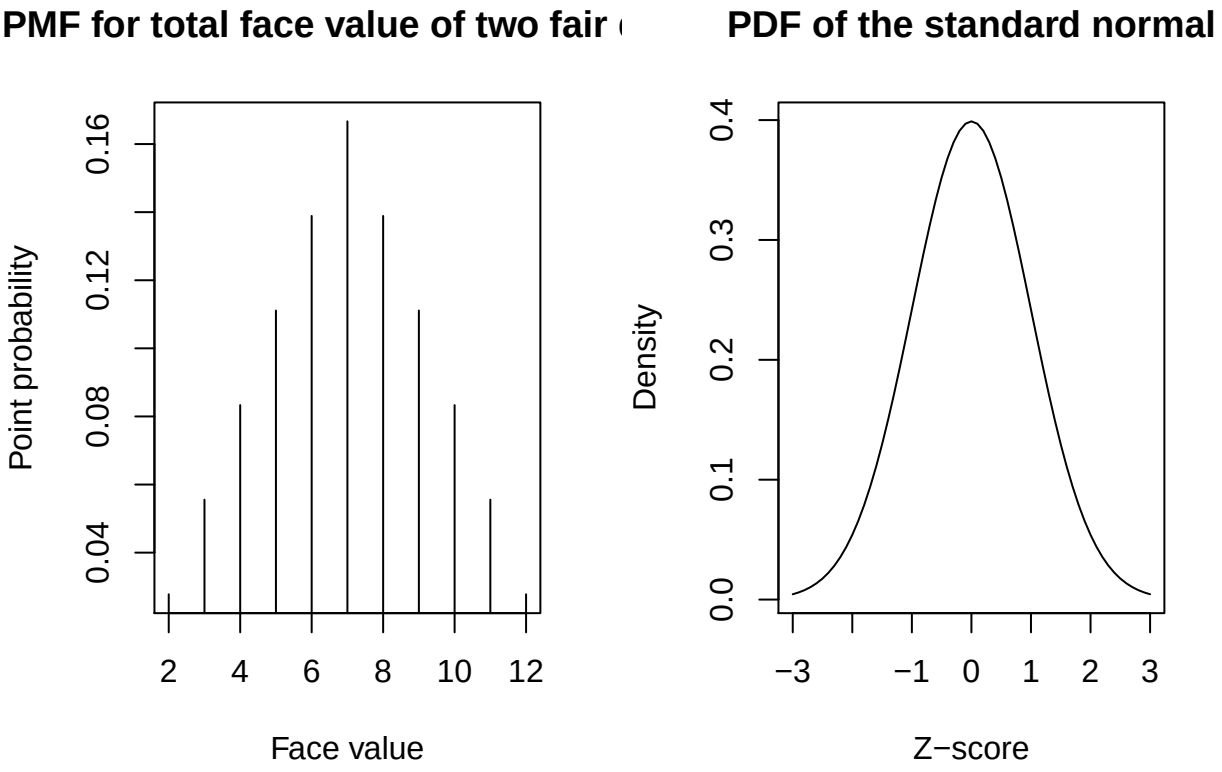
```
  # initialize an empty vector of length 12 with all zeroes
  combos_2dice = rep(0,12)
  for (i in 1:6) {  # compute all the combinations
    for (j in 1:6) {
      index = i+j
      combos_2dice[index] = combos_2dice[index] + 1
    }
  }
  return(combos_2dice[value]) # return the value(s) of interest
}
face.count = ncomb.2dice(2:12)
two_dice = data.frame(sum = 2:12, count = face.count, prop = face.count/sum(face.count))


# ============================================================================ #
# standard normal
z.scores = seq(-3,3,by=.1)
d.values = dnorm(z.scores)

# ============================================================================ #
# plots
par(mfrow=c(1,2))

plot(two_dice$sum, two_dice$prop, type="h",
     xlab="Face value", ylab="Point probability",
     main = "PMF for total face value of two fair dice")

plot(z.scores, d.values,
     type = "l",
     xlab="Z-score", ylab="Density",
     main = "PDF of the standard normal")
```



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.)

⇒ ***Question: ?***

Answer

- Answer.