Probability

Dr. Michael Fix
mfix@gsu.edu

Georgia State University

18 January 2024

Introducing Probability

- Probability is the limiting factor of relative frequency
- Probability is indicated as number between 0 (never happens) and 1 (always happens)
- In a frequency table:

	Relative Frequency		
Outcome	10 Tosses	50 Tosses	∞ Tosses
Heads	0.3	0.54	0.5
Tails	0.7	0.46	0.5
Total	1.0	1.0	1.0

- In mathematical symbols:

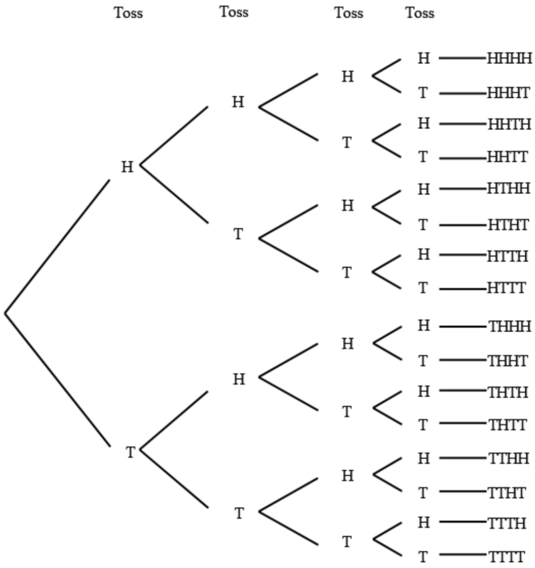
$$Pr \equiv \lim(\frac{f}{n}) \tag{1}$$

Basics of Probability Models

- These core principles are useful to us in a number of ways
- One is to work backwards and use samples to estimate population distributions
- The other is to use known (population) probabilities to determine the likelihood of a particular event occurring
  - Likelihood is simple with something like a single coin toss: each toss is equally likely to be heads or tails
  - Probability gets more complicated when we start to consider multiple events: what is the probability of throwing heads 3 time in a row?

Probability Models

Tree Model



Probability Models

Tree Model

- This model allows you to see the possible outcomes
- For three coin tosses:
  - After one toss there are two outcomes: heads or tails ($Pr_{heads} = 0.5$)
  - After two tosses there are now four outcomes: heads or tails following the initial heads or tails ($Pr_{2heads} = 0.25$)
  - After three there are now eight outcomes in the same pattern: ($Pr_{3heads} = 0.125$)
- All together these eight outcomes are known as the outcome set (S)

Combining Probabilities

- We can generalize this logic:
- The previous example asked the likelihood of heads *and* heads *and* heads
- For cases of *and* ($\cap$) we multiply probabilities
- Probabilities range from 0 to 1, so multiplying fractions leads to smaller fractions (less likely events)
- This is known as a **joint probability**
- What about something slightly more complex: the probability of at least two heads in three flips?

Combining Probabilities

- We now have more than one way to get an outcome:
- Head Head Head *or* Head Tail Head *or* Tail Head Head *or* Tail Head Tail
- For cases of *or* ($\cup$) we add probabilities:
- $Pr_{2heads} = 0.125 + 0.125 + 0.125 + 0.125 = 0.5$
- The probability of the combined events (E) is a subset of the full outcome set (S)
- and is the sum of the constituent events $(e_1, e_2, \ldots)$
- Formally:

$$Pr(E) = \sum Pr(e) \quad (2)$$

Combining Subsets

- Do the subsets (E and F for example) overlap?
  - If not, combining subsets is as simple as single events ($E + F$)
  - If so, combinations need to account for overlap
- Combined with *or*: count all outcomes in E, F, or both sets
- Combined with *and*: count just those in both sets
- Combining two overlapping sets: add the *or* and subtract the *and*

$$Pr(E \cup F) = Pr(E) + Pr(F) - Pr(E \cap F) \tag{3}$$

Combining Subsets

- Mutual exclusivity exists when there is no overlap between subsets: $Pr(E \cap F) = \emptyset$
- So:

$$Pr(E \cup F) = Pr(E) + Pr(F) \tag{4}$$

- Each subset also has an opposite set known as a complement
  - The complement ($\bar{E}$) is every outcome *not* contained in the set (E)
  - For example, the complement of throwing at least two heads is the probability of throwing fewer than two heads

$$Pr(E) = 1 - Pr(\bar{E}) \tag{5}$$

Conditional Probability

- Conditional Probability involves calculating probability based on a limiting condition
- We are looking for the probability of outcome (E) given the condition (F)
- For example: the probability of tossing at least two heads given the first toss was heads
- Formally:

$$Pr(E|F) = \frac{Pr(E \cap F)}{Pr(F)} \tag{6}$$

- With a little algebra we can solve for $Pr(E \cap F)$

$$Pr(E \cap F) = Pr(F)Pr(E|F) \tag{7}$$

Statistical Independence

- A special case of conditional probability
- Unconditional probability of an event is the same as the conditional probability
- Indicates that the condition has no impact on the probability
- Formally:

$$Pr(F|E) = Pr(F) \tag{8}$$

- Statistical independence works both ways: if E is independent of F then F is independent of E.
- Independence simplifies combined probabilities:

$$Pr(E \cap F) = Pr(E)Pr(F) \tag{9}$$

Bayes Theorem

- The same problem looked at from a different perspective
- Reverses the probability tree
- Two key terms:
  - Prior Probabilities: initial (hypothesized) probabilities before testing
  - Posterior Probabilities: probabilities (observed) after testing

$$Pr(E|F) = \frac{Pr(F|E)Pr(E)}{Pr(F)} \tag{10}$$

Introducing Probability Distributions

- Simple probability tables and trees work for small events
- How do we handle probabilities in larger data?
- Probability distributions represent these more effectively
- Different types of data utilize different types of probability distributions
  - Discrete variables
  - Continuous variables

Discrete Probability Distributions

- Discrete probability distributions are based on frequency
- A discrete random variable can only take on specific values (0 or 1, count data, etc.)
- The probability distribution of a discrete random variable determines the frequency of a random sample
- Random samples of the population should provide an approximation of the population parameters
- Probability distributions can also provide descriptive information about the shape of the data (moments)
- Probability distributions describe the population distribution

Discrete Probability Distributions

Sample and Population Values

- Recall that sample statistics are indicated by Latin letters
- Mean and Variance can be derived from the probability distribution $p(x)$ in addition to relative frequency $(\frac{f}{n})$
- Mean of sample:

$$\bar{X} = \sum x(\frac{f}{n}) \tag{11}$$

- Variance of sample:

$$s^2 = \sum (x - \bar{X})^2 (\frac{f}{n}) \tag{12}$$

Discrete Probability Distributions

Sample and Population Values

- Recall that population parameters by Greek letters
- Mean of Population:

$$\mu \equiv \sum x p(x) \tag{13}$$

- Variance of Population:

$$\sigma^2 \equiv \sum (x - \mu)^2 p(x) \tag{14}$$

- Or more simply:

$$\sigma^2 = \sum x^2 p(x) - \mu^2 \tag{15}$$

The Binomial Distribution

Definitions and Assumptions

- A specific form of a discrete probability distribution
- Discrete values can only be one of two values (hence: *bi*nomial)
- Useful for measuring events: did it occur or not?
- Notation and assumptions:
  - *n* - number of observations in the sample
  - *s* - number of successes in the sample
  - *s* occurs with probability π, the opposite is the compliment of π $(1 - \pi)$
  - Assume that each observation is statistically independent

The Binomial Distribution

Formula and Notation

- The general form:

$$p(s) = \binom{n}{s} \pi^s (1-\pi)^{n-s} \tag{16}$$

- The binomial coefficient $\binom{n}{s}$ is defined as:

$$\binom{n}{s} = \frac{n!}{s!(n-s)!} \tag{17}$$

The Binomial Distribution

Formula and Notation

- The general form:

$$p(s) = \binom{n}{s} \pi^s (1-\pi)^{n-s} \tag{16}$$

- The binomial coefficient $\binom{n}{s}$ is defined as:

$$\binom{n}{s} = \frac{n!}{s!(n-s)!} \tag{17}$$

- And finally n factorial ($n!$) is:

$$n! = n(n-1)(n-2)\cdots 1 \tag{18}$$

The Binomial Distribution

Uses of the Distribution

- Sampling from large populations
  - In sufficiently large samples, random draws can adequately satisfy independence assumption without the need for replacement
  - The binomial distribution can help us determine the likelihood of drawing a certain number of successes (or failures)
  - This provides a basis for measures of confidence in causal analysis

Continuous Distributions

General Concepts

- Not all data is discrete, many measures have an infinite number of possible observations
- Simple frequencies do not work with continuous data without blocking observations (1-10, 10-20, etc.)
- Frequency can be shown as blocks in a histogram, as the number of observations increase and block size decreases we can represent frequency as a density curve
- The probability distribution then is measured by the area under a range of the curve rather than the size of the block (as in discrete data)

Continuous Distributions

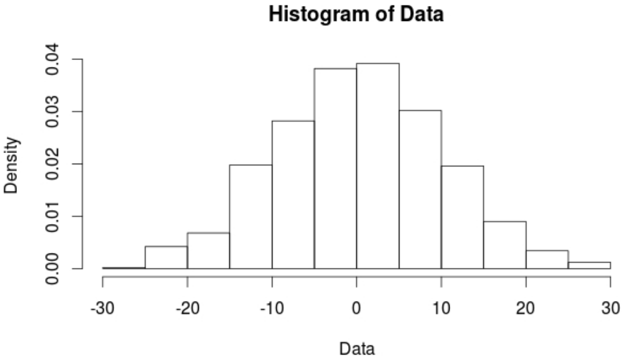
General Concepts

- The probability distribution then is measured by the area under a range of the curve rather than the size of the block (as in discrete data)
- Measuring area under curves means using integral calculus to calculate probability
- Using relative frequency density is more useful than raw frequency
- In this case the sum of the area within the histogram bars (or area under curve) is equal to 1

$$Relative\ frequency\ density\ \equiv \frac{relative\ frequency}{cell\ width} \tag{19}$$

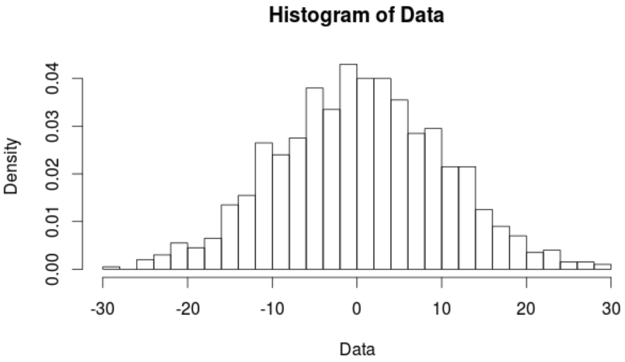
Visualizing A Continuous Distribution

- Start with a typical histogram with the observations in around 10 blocks



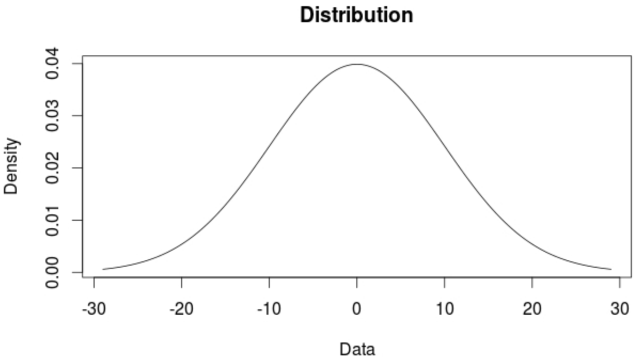
Visualizing A Continuous Distribution

- Double the number of blocks



Visualizing A Continuous Distribution

- Finally smooth it out to infinite observations
- This is the underlying probability distribution



The Normal (Z) Distribution

- This is the standard "bell curve"
- Also known as the Z distribution or Gaussian curve
- The standard normal distribution
  - $\mu = 0$
  - $\sigma = 1$
  - Z is the number of standard deviations a given point is from the mean
- The area under the curve represents the probability of a value
- Above (or below) a certain Z value the area provides the probability of a value higher (or lower) than that value

The General Normal Distribution

- Not all normal distributions are centered on 0 with $\sigma = 1$
- The distribution can be generalized to other values of μ and σ
- Z remains the measure of distance from the center in standard deviations
- Z can be calculated for observation X as:

$$Z = \frac{X - \mu}{\sigma} \tag{20}$$

Distributions with Two Variables

- Combining probability distributions to work with two variables
- Using more than one variable allows us to do more than describe a distribution
- Builds on concepts of probability and probability distributions
- Introducing new notations for effectively handling two variables

Distributions with Two Variables

Joint Distributions

- Joint distributions allow for showing probability relationships with two variables
- The goal is to show what the probability of a given X value (x) *and* a given Y value (y)
- In familiar notation:

$$Pr(X = x \cup Y = y) \tag{21}$$

- Becomes:

$$p(x, y) \tag{22}$$

Marginal Distributions

- Marginal distributions focus on the distribution of one variable given the other is held constant
- For example: distribution of x given $y = 2$
- Formally this is represented as:

$$p(x) = \sum_{y} p(x, y) \tag{23}$$

- Intuitively, we can see the sums in the margins of a table

x	y = 1	2	$p(x)$
1	0.1	0.4	0.5
2	0.3	0.2	0.5
$p(y)$	0.4	0.6	1.00

Independence for Joint Probability Distributions

- Independence for all values x and y is denoted as:

$$p(x, y) = p(x)p(y) \tag{24}$$

- This means the product of the marginal values for each cell must equal the value in the cell
- Tables of independent joint distributions must be proportional in both the columns and the rows

	y		
x	1	2	$p(x)$
1	0.2	0.3	0.5
2	0.2	0.3	0.5
$p(y)$	0.4	0.6	1.00