5. Introduction to Random Variables

Terminology[edit] As probability theory is used in quite diverse applications, terminology is not uniform and sometimes confusing. The following terms are used for non-cumulative probability distribution functions:

Probability mass, Probability mass function, p.m.f.: for discrete random variables. Categorical distribution: for discrete random variables with a finite set of values. Probability density, Probability density function, p.d.f.: most often reserved for continuous random variables. The following terms are somewhat ambiguous as they can refer to non-cumulative or cumulative distributions, depending on authors' preferences:

Probability distribution function: continuous or discrete, non-cumulative or cumulative. Probability function: even more ambiguous, can mean any of the above or other things. Finally,

Probability distribution: sometimes the same as probability distribution function, but usually refers to the more complete assignment of probabilities to all measurable subsets of outcomes, not just to specific outcomes or ranges of outcomes.

1 What Is a Probability Distribution?

- If you spend much time at all dealing with statistics, pretty soon you run into the phrase probability distribution.
- It is here that we really get to see how much the areas of probability and statistics overlap.
- Although this may sound like something technical, the phrase probability distribution is really just a way to talk about organizing a list of probabilities.
- A probability distribution is a function or rule that assigns probabilities to each value of a random variable. The distribution may in some cases be listed. In other cases it is presented as a graph.

2 What is a Probability Distribution

A statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. This range will be between the minimum and maximum statistically possible values, but where the possible value is likely to be plotted on the probability distribution depends on a number of factors, including the distributions mean, standard deviation, skewness and kurtosis.

Thirty-eight students took the test. The X-axis shows various intervals of scores (the interval labeled 35 includes any score from 32.5 to 37.5). The Y-axis shows the number of students scoring in the interval or below the interval.

cumulative frequency distributionA can show either the actual frequencies at or below each interval (as shown here) or the percentage of the scores at or below each interval. The plot can be a histogram as shown here or a polygon.

Probability Distributions (Question 2 for End Of Year Exam)

- Discrete Probability Distributions
  - Binomial Probability Distribution (Week 3)

- Geometric Probability Distribution (Week 3)
- Poisson Probability Distribution (Week 3/4)

- Continuous Probability Distributions
  - Exponential Probability Distribution (Week 4)
  - Uniform Probability Distribution (Week 4)
  - Normal Probability Distribution (Week 4/5)

2.1 Types of Probability Distribution

- A probability distribution is a mathematical approach to quantifying uncertainty.
- There are two main classes of probability distributions: Discrete and continuous.
  - Discrete distributions describe variables that take on discrete values only (typically the positive integers),
  - continuous distributions describe variables that can take on arbitrary values in a continuum (typically the real numbers).

3 Probability Distributions

Probability Distributions will covered in detail in this course as part of the `R` component (i.e. weeks 8-13).

It is worth bearing in mind that all of the material that will be covered in the `R` section of the course can just as easily be implemented using MATLAB. In fact substantial use is made of these commands in real world applications, particularly in Finance and Engineering. As such we will briefly look at some.

- Suppose we have a set of **n** items.
- From that set, we create a subset of **k** items.
- The **order** in which items are selected is recorded. (The ordering of selected items is very important.)
- The total number of **ordered subsets** of **k** items chosen from a set of **n** items is

$$\frac{n!}{n-k!}$$

4 Quantiles for Probability Distributions

- The quantile (this term was first used by Kendall, 1940) of a distribution of values is a number xp such that a proportion p of the population values are less than or equal to xp.

- For example, the .25 quantile (also referred to as the 25th percentile or lower quartile) of a variable is a value (xp) such that 25% (p) of the values of the variable fall below that value.
- Similarly, the 0.75 quantile (also referred to as the 75th percentile or upper quartile) is a value such that 75% of the values of the variable fall below that value and is calculated accordingly.

Joint probability tables

A joint probability table is a table in which all possible events (or outcomes) for one variable are listed as row headings, all possible events for a second variable are listed as column headings, and the value entered in each cell of the table is the probability of each joint occurrence.

Often, the probabilities in such a table are based on observed frequencies of occurrence for the various joint events. The table of joint-occurrence frequencies which can serve as the basis for constructing a joint probability table is called a contingency table.

1. A pair of dice is thrown. Let X denote the minimum of the two numbers which occur. Find the distributions and expected value of X.
2. A fair coin is tossed four times. Let X denote the longest string of heads. Find the distribution and expectation of X.
3. A fair coin is tossed until a head or five tails occurs. Find the expected number E of tosses of the coin.
4. A coin is weighted so that $P(H) = 0.75$ and $P(T) = 0.25$

   The coin is tossed three times. Let X denote the number of heads that appear.

   - (a) Find the distribution f of X.
   - (b) Find the expectation E(X).

4.1 Graph of a Probability Distribution

A probability distribution can be graphed, and sometimes this helps to show us features of the distribution that were not apparent from just reading the list of probabilities. The random variable is plotted along the x-axis, and the corresponding probability is plotted along the y - axis.

- For a discrete random variable, we will have a histogram
- For a continuous random variable, we will have the inside of a smooth curve

The rules of probability are still in effect, and they manifest themselves in a few ways. Since probabilities are greater than or equal to zero, the graph of a probability distribution must have y-coordinates that are nonnegative. Another feature of probabilities, namely that one is the maximum that the probability of an event can be, shows up in another way.

$$\text{Area} = \text{Probability}$$