ST1050

Class 7

- Distributions

Learning Objectives

1. Define "distribution"
2. The difference between a discrete and continuous distribution
3. Distinguish between a frequency distribution and a probability distribution
4. Recognise the Normal, Binomial and Uniform distributions

Distributions

The *distribution* of a statistical data set (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur.

If a variable can take on any value between its minimum value and its maximum value, it is called a continuous variable; otherwise, it is called a categorical or discrete variable.

Some examples will clarify the difference between discrete and continuous variables:

- Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.

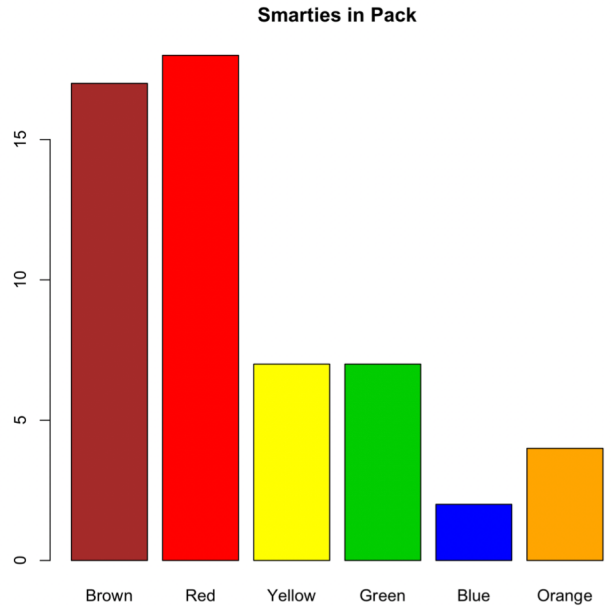
- Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

When a distribution of categorical or discrete data is organized, you see the number or percentage of individuals in each group. When a distribution of continuous data is organized, they're often ordered from smallest to largest, broken into reasonably sized groups (if appropriate), and then put into graphs and charts to study the shape, centre, and amount of variability in the data.

Distributions of Categorical/Discrete Variables:

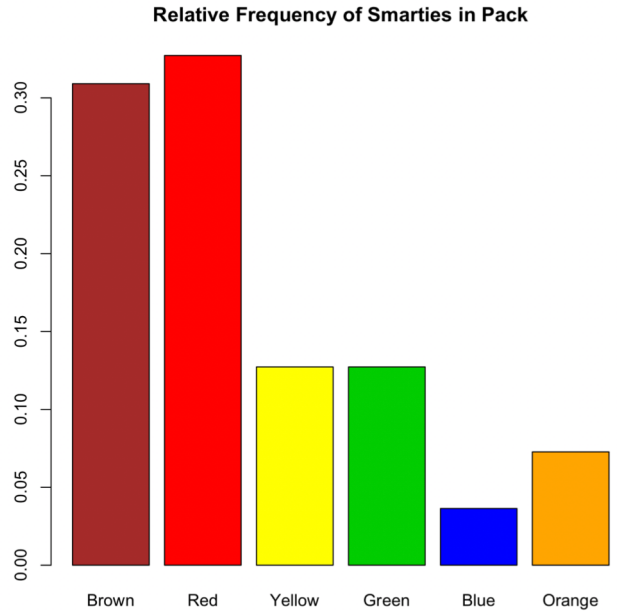
Imagine you purchase a pack of Smarties.
The Smarties come in six different colours. A quick count showed that there were 55 Smarties distributed as follows:

Colour	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4



This table is called a frequency table and it describes the distribution of SMARTIES colour frequencies. Not surprisingly, this kind of distribution is called a frequency distribution. Often a frequency distribution is shown graphically. To the right is a barplot showing the distribution. If you divide by 55 the numbers plotted in the barplot are the relative frequencies.

Note: this was made using the barplot() function – it is not a histogram. We will look at barplots more later on.

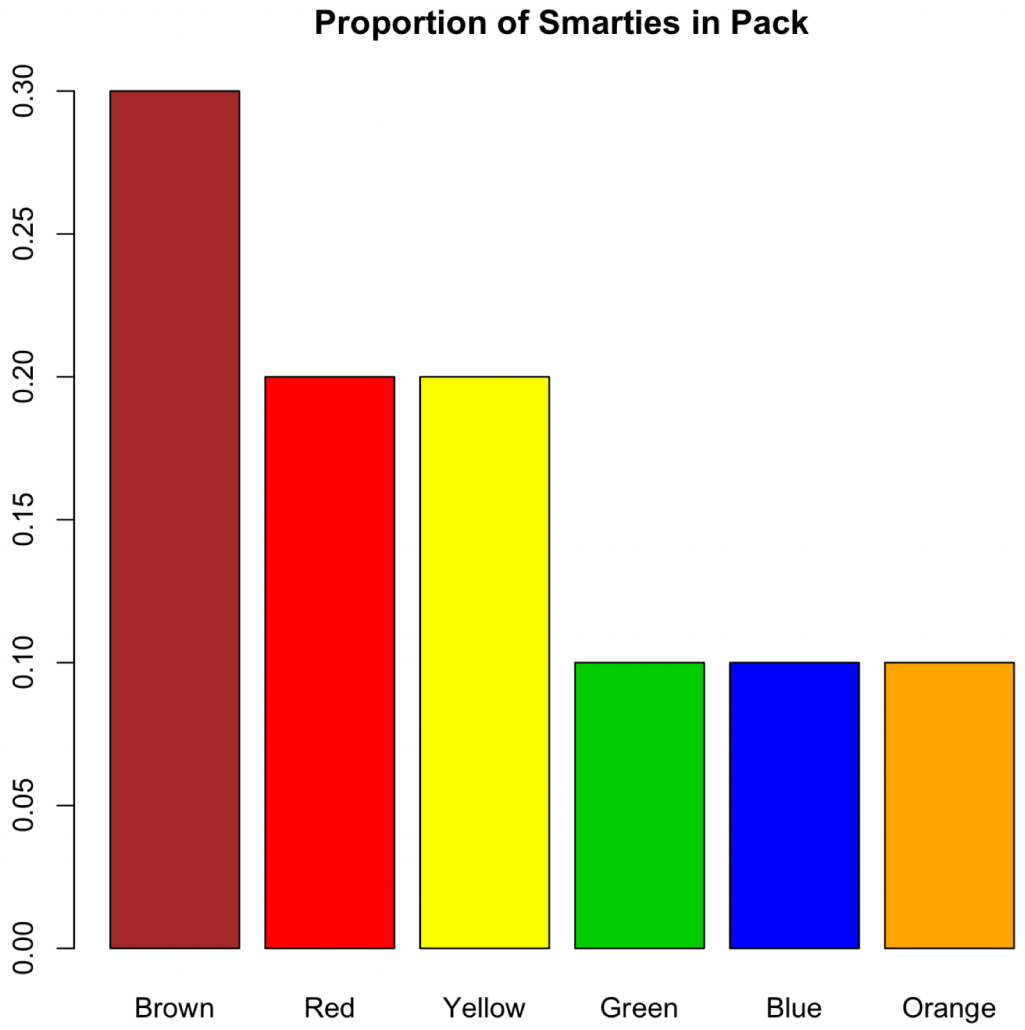


Distributions of Categorical/Discrete Variables (cont'd):

The distribution shown on the previous slide concerns just one bag of Smarties. You might be wondering about the distribution of colours for all Smarties.

The manufacturer of Smarties provides some information about this matter, but they do not tell us exactly how many Smarties of each colour they have ever produced. Instead, they report proportions rather than frequencies. The barplot to the right shows these proportions.

Since every Smartie is one of the six colours, the six proportions shown in the figure add to one.
We call the information in this barplot a probability distribution because if you choose a Smartie at random, the probability of getting, say, a brown Smartie is equal to the proportion of Smarties that are brown (0.30).



Distributions of Categorical/Discrete Variables (cont'd):

Notice that the distributions in the two barplots are not identical.

The first plots portray the distribution in a sample of 55 Smarties; they are a frequency and relative frequency distribution. The second plot shows the proportions for all Smarties; it is a probability distribution.

Chance factors involving the machines used by the manufacturer introduce random variation into the different bags produced. Some bags will have a distribution of colours that is close to the probability distribution; others will be further away.

One could say that the relative frequency distribution is a data-dependent proxy for the underlying probability distribution.

Continuous Distributions:

Examples:

- Time it takes a computer to complete a task. You might think you *can* count it, but time is often rounded up to convenient intervals, like seconds or milliseconds. Time is actually a continuum: it could take 1.3 seconds or it could take 1.333333333333333… seconds.
- A person's weight. Someone could weigh 180 pounds, they could weigh 180.10 pounds or they could weigh 180.1110 pounds. The number of possibilities for weight are limitless.
- Income. You might think that income is countable (because it's in dollars) but who is to say someone can't have an income of a billion dollars a year? Two billion? Fifty nine trillion? And so on…
- Age. So, you're 25 years-old. Are you sure? How about 25 years, 19 days and a millisecond or two? Like time, age can take on an infinite number of possibilities and so it's a continuous variable.
- The price of gas. Sure, it might be $4 a gallon. But one time in recent history it was 99 cents. And give inflation a few years it will be $99. not to mention the gas stations always like to use fractions (i.e. gas is rarely $4.47 a gallon, you'll see in the small print it's actually $4.47 9/10ths

The Normal Distribution:

Distributions for continuous variables are called continuous distributions. (They also can be called probability densities). Some continuous distributions have particular importance in statistics. A very important one is shaped like a 'bell', and called the normal distribution. Many naturally-occurring phenomena can be approximated surprisingly well by this distribution. It will serve to illustrate some features of all continuous distributions.

Some Properties of a normal distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the centre (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

Examples:

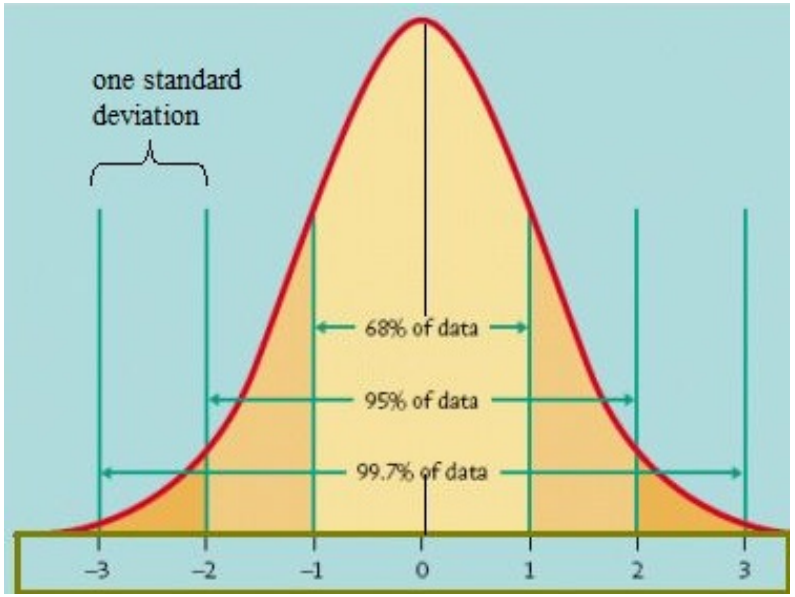
- Heights of people.
- Measurement errors.
- Blood pressure.
- Points on a test.
- IQ scores.
- Salaries.

The Normal Distribution (cont'd):

Standard Normal Model: Distribution of Data

The normal distribution is often called a bell curve: it has a small percentage of the points on both tails and the bigger percentage on the inner part of the curve.

In the **standard normal model**, about 5 percent of your data would fall into the “tails” (coloured darker orange in the image below) and 90 percent will be in between. For example, for test scores of students, the normal distribution would show 2.5 percent of students getting *very* low scores and 2.5 percent getting *very* high scores. The rest will be in the middle; not too high or too low.



Normal distribution functions in R

R has functions for obtaining probabilities, quantiles and random samples from various distributions.
Let's start with the Normal distribution.
By default R will use the standard Normal distribution N(0,1). But you can specify other values for the mean and variance if necessary:

`rnorm(10)`	random sample of 10 from the standard normal distribution
`pnorm(1.7)`	Probability of a value less than 1.7 in the standard normal distribution
`qnorm(0.85)`	85^{th} percentile of the standard normal distribution
`help(pnorm)`	information on the pnorm function

Example:
Suppose that the price of 5 year-old used cars is known to follow a Normal distribution with mean €4000 and sd 500. A particular car is selling at €5500. Would you conclude this to be an abnormal price?
Find the relative frequency of values at or above 5500 when the population distribution is normal with mean 4000 and sd 500.

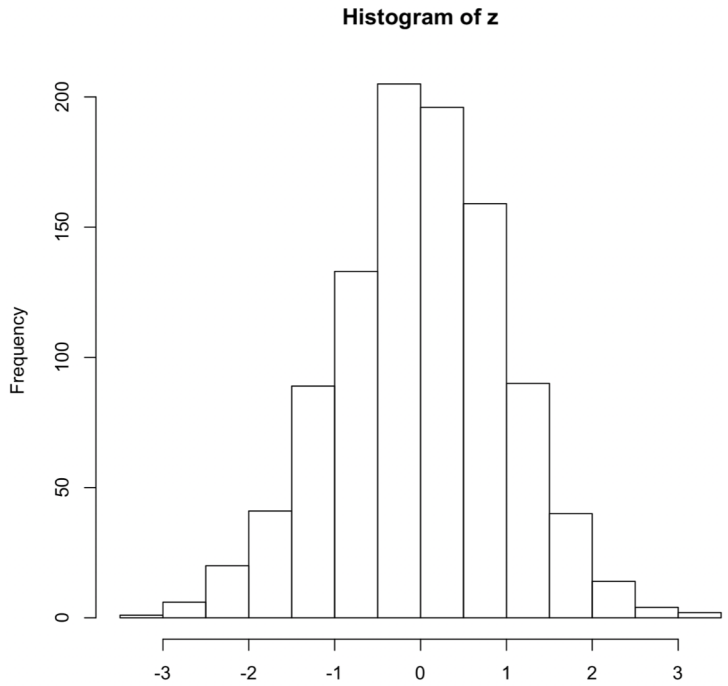
`pnorm(5500,mean=4000,sd=500)` gives the relative frequency of values less than 5500.
We would like to know the relative frequency of values greater than 5500, so we look at:

```
1-pnorm(5500,4000,500)
```

This value is very small hence we may not want to buy this car!!

The Normal Distribution (cont'd):

This is a histogram created from z= rnorm(1000,0,1) in R.



Example:
Assume a random variable Z is distributed according to the normal distribution with mean 6 and standard deviation 4.
What is the probability that Z takes on a value between -1 and 3 ?

(hint: subtract the c.d. at -1 from the c.d. at 3)
pnorm(3, 6, 4) - pnorm(-1, 6, 4)
[1] 0.1865682

The Binomial Distribution:

The ***binomial distribution*** describes the behaviour of a count variable *X* if the following conditions apply:

1: *The number of observations n is fixed.*
2: *Each observation is independent.*
3: *Each observation represents one of two outcomes ("success" or "failure").*
4: *The probability of "success" p is the same for each outcome.*

If these conditions are met, then *X* has a binomial distribution with parameters *n* and *p*, abbreviated *B(n,p)*.

The Binomial Distribution (cont'd):

The binomial distribution model deals with finding the probability of success of an event which has only two possible outcomes in a series of experiments. For example, tossing of a coin always gives a head or a tail. The probability of finding exactly 3 heads in tossing a coin repeatedly for 10 times is estimated using the binomial distribution.

R has four in-built functions to generate binomial distribution.

dbinom(x, size, prob)
pbinom(x, size, prob)
qbinom(p, size, prob)
rbinom(n, size, prob)

This is the description of the parameters used −
x is a vector of numbers.
p is a vector of probabilities.
n is number of observations.
size is the number of trials.
prob is the probability of success of each trial.

The Binomial Distribution (cont'd):

pbinom()
This function gives the cumulative probability of an event. It is a single value representing the probability.
e.g. Probability of getting 26 or less heads from a 51 tosses of a coin.

```
x <- pbinom(26,51,0.5); print(x)
[1] 0.610116
```

rbinom()
This function generates required number of random values of given probability from a given sample.
In the rbinom function, n is the length of the vector to create, while size is the number of items to hypothetically draw from a theoretical urn having a distribution determined by 'prob'. The result returned is the number of 'ones' that you draw.
eg: Find 8 random values from a sample of 150 with probability of 0.4.

```
x <- rbinom(8,150,.4);  print(x)
[1] 49 44 55 52 53 55 62 62
```

The Binomial Distribution (cont'd):

Examples

Q: Assume a coin is weighted so that it comes up heads 60% of the time. What is the probability that you will obtain 25 or more heads after 50 flips?

Hint: Use pbinom to get the probability of 25 or less heads, and subtract from 1

```
1 - pbinom(25,50,0.6)
[1] 0.9021926
```

Q: Assume a standard die is rolled 10 times. What it the probability that you will roll fewer than 5 sixes?

Hint: Use pbinom to sum the cases 0, 1, 2, 3, and 4 and subtract from 1.

```
pbinom(4, 10, 0.16)
[1] 0.9869899
```

Uniform distribution

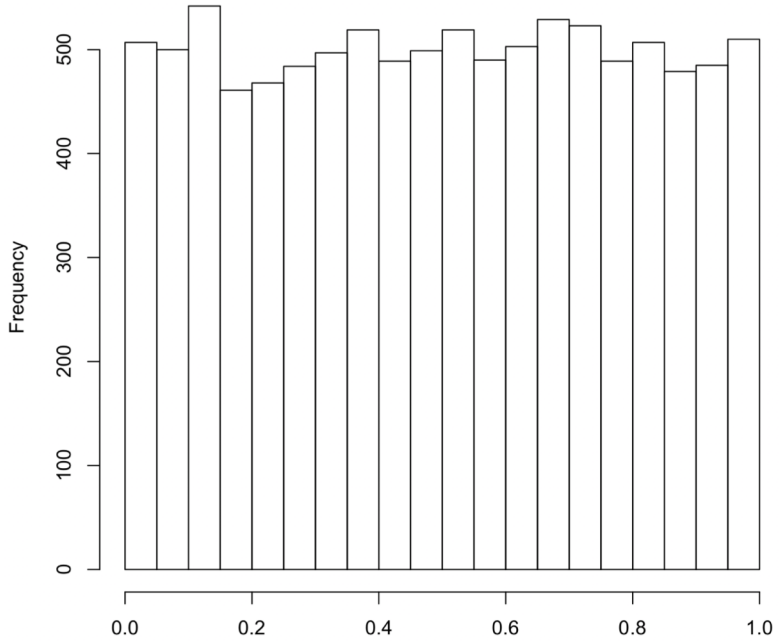
Uniform distribution, is a distribution in which every possible result is equally likely; that is, the probability of any individual event occurring is the same.

Consider the toss of a single die. The outcome of this toss is a random variable that can take on any of six possible values: 1, 2, 3, 4, 5, or 6. Each of these outcomes is equally likely to occur.

The probability that any particular outcome will occur is equal to 1/6. Therefore, the outcome from the toss of a single die has a uniform distribution.

In R the Uniform distribution functions are:
dunif(x, min = 0, max = 1, log = FALSE)
punif(q, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE)
qunif(p, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE)
runif(n, min = 0, max = 1)

for example, if we create data as z = runif(10000) this is what the histogram is like:



Poisson Distribution

A Poisson distribution is the probability distribution that results from a Poisson experiment.

Attributes of a Poisson Experiment

A **Poisson experiment** is a statistical experiment that has the following properties:

- The experiment results in outcomes that can be classified as successes or failures.
- The average number of successes (μ) that occurs in a specified region is known.
- The probability that a success will occur is proportional to the size of the region.
- The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

The Poisson distribution has the following properties:

The mean of the distribution is equal to μ .

The variance is also equal to μ .

Practical Applications of A Poisson Distribution

The Poisson distribution is commonly used within industry and the sciences.

eg:

- number of equipment failures per day for logistics company
- number of customers arriving at a retailer
- the number of visitors to a web site
- number of inbound phone calls
- number of customer complaints

Example:

Assume a a ball from the driving range next door lands in your yard at an average rate of 3 balls per hour during the day. What is the probability that 10 or fewer golf balls will land in your yard during the afternoon, assuming the afternoon is 5 hours long?

Hint: mean is 15 = 3 * 5 for the entire afternoon ppois(10, 15)

```
1-ppois(10, 15)
[1] 0.2211992
```