Section 4.3 - Random Variables

Statistics 104

Autumn 2004

Random Variables

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon

Examples:

- Sum of rolling two 4-sided dice
- Number of faulty switches out of 6 randomly drawn from a batch
- Throw a dart at a dart board. Measure the distance from the center.
- Time you arrive in the classroom

The first two are examples of **discrete** random variables and the last two are examples of **continuous** random variables (RV).

Want to talk about models for these two types of random variables.

Discrete Random Variables

A discrete random variable X takes a discrete set of possible values. The probability distribution of X lists the possible values and their corresponding probabilities

Value of X	x_1	x_2	x_3	$\ldots$	x_k
Probability	p_1	p_2	p_3	$\ldots$	p_k

where $p_j = P[X = x_j], j = 1, \ldots, k$

These probabilities must satisfy

- $0 \leq p_j \leq 1$ for each j
- $p_1 + p_2 + \ldots + p_k = 1$

Note that the book says that k, the number of different possible values for X, should be finite. However it is possible to have an infinite number of possibilities.

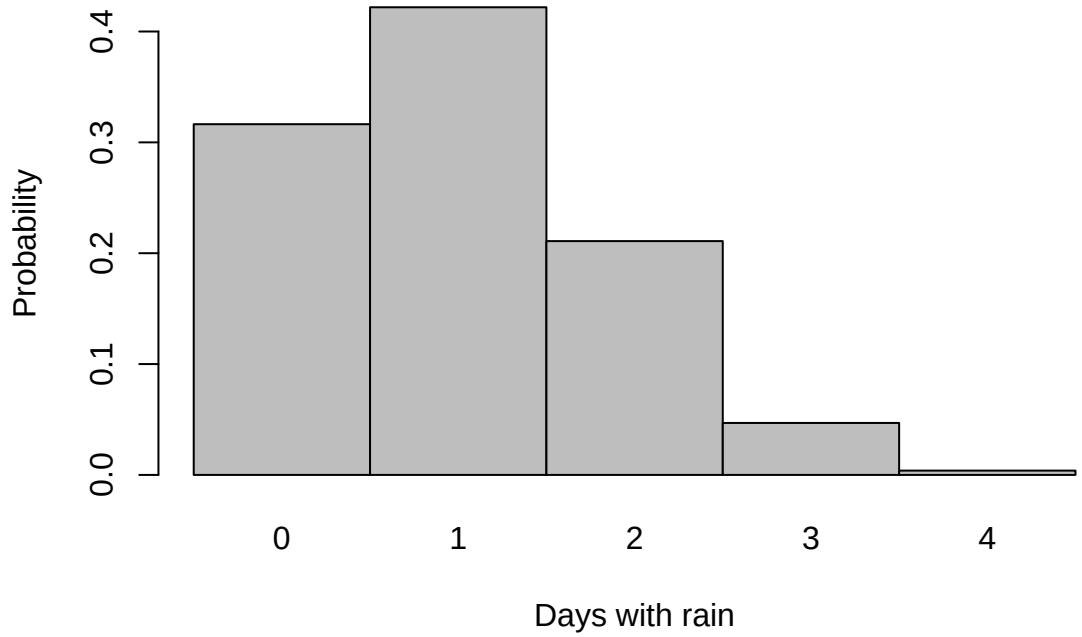
$P[X \text{ in } A]$ is found by summing the p_j's for the x_j's in A.

Example: Suppose that for each of the next 4 days, the chance of rain on each day is 0.25. The probability distribution for the number of days with recordable rain, assuming days are independent (a very dubious assumption), is

x_j	$p_j = P[X = x_j]$
0	0.3164
1	0.4219
2	0.2109
3	0.0469
4	0.0039

Probability Histograms

An approach for displaying discrete probability distributions. Similar to histograms of data, except that instead of plotting the number (or proportion) of observations in each class, the heights of the bars are the probabilities for each possible outcomes.



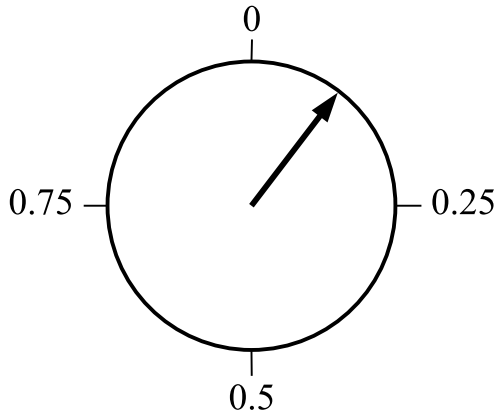
Have one bar per possible outcome.

Continuous Random Variables

Continuous random variables (potentially) take on any possible value in some range.

Examples:

- Distance from bullseye on a dart board
- People's heights and weights
- Position of a spinner



- Deviation from scheduled arrival time for flights arriving at Logan
- Stock returns, change in share prices, etc

Note that strictly, variables like share prices are discrete, but since they are on such a fine scale, they are usually considered as continuous.

The ranges for continuous RVs may, or may not, be bounded.

The spinner can only take values in $[0, 1)$, distances from the bullseye take values in $[0, \infty)$, and deviations from scheduled arrival times take values in $(-\infty, \infty)$.

We are interested in probabilities like $P[X > \frac{1}{2}]$ or $P[\frac{1}{3} < X < \frac{2}{3}]$.

We can't take the approach used for discrete random variable by assigning a probability to each possible x as there uncountably infinite number of points.

Instead, probabilities are based on the idea of a density curve (see section 1.3)

Density Curves

A density curve must satisfy the following two conditions

1. Is always on or above the horizontal axis $\{f(x) \geq 0\}$

2. Has area exactly 1 underneath it

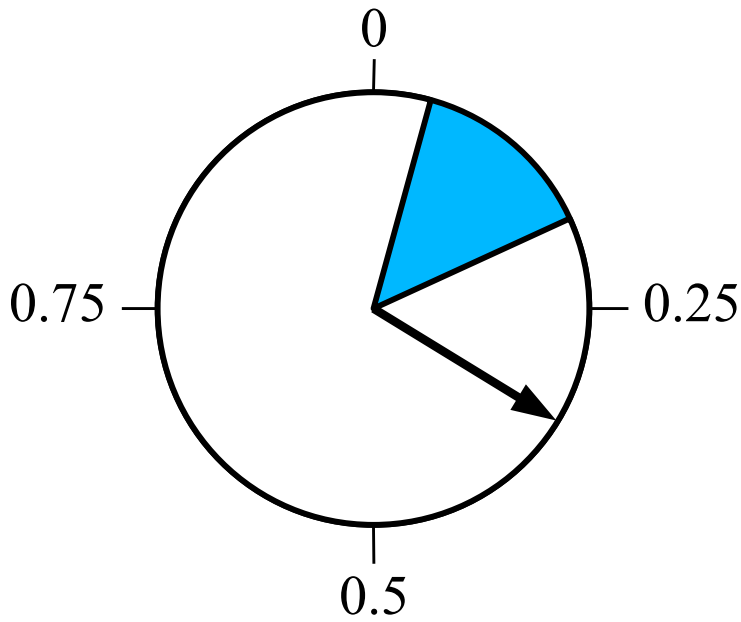
The height of the curve describes the likelihood of getting each possible outcome,

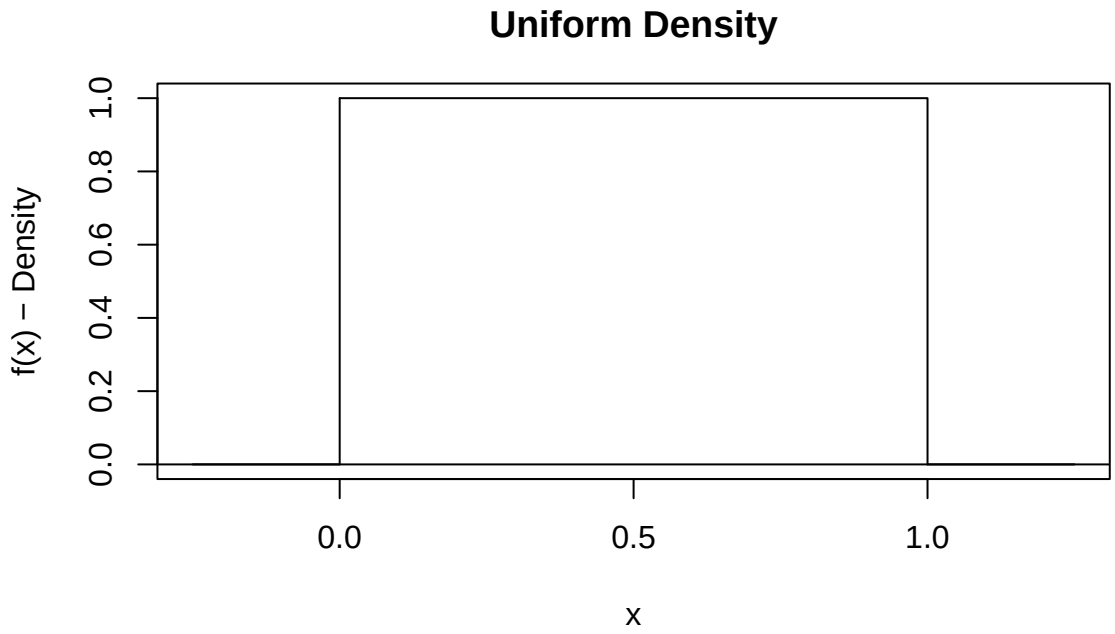
Areas under the curve give probabilities.

- Condition 1 is the analogue to $0 \leq p_j \leq 1$ for discrete random variables

- Condition 2 is the analogue to $\sum_{j=1}^{k} p_j = 1$ for discrete random variables. It is a restatement of $P[S] = 1$ from the general rules of probability.

Spinner example: Lets assume that every direction on the spinner is equally likely. This suggests that the height of the curve should be the same for all possible values

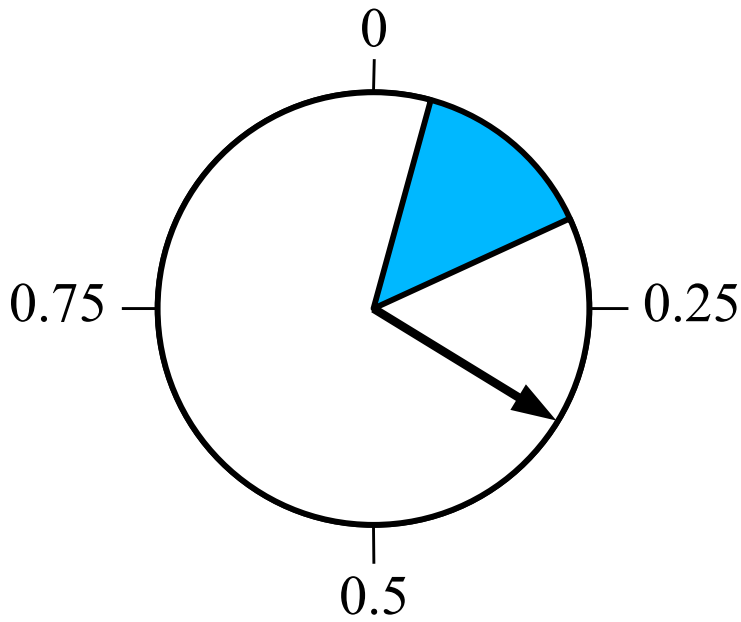


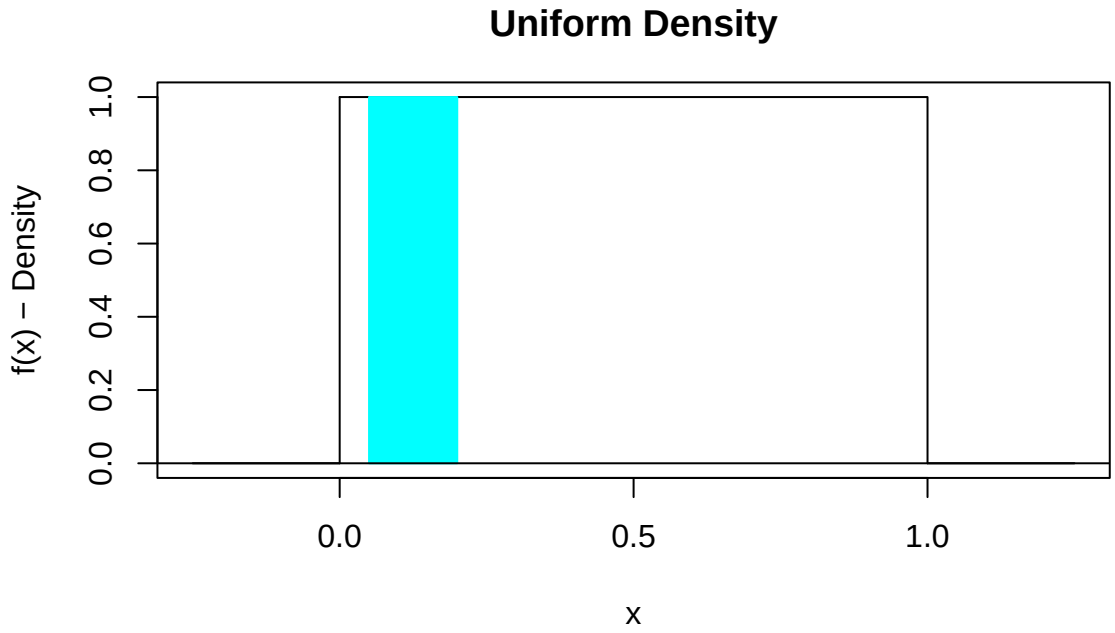


With this uniform density, the probability of any event of the form $A = [a, b]$ is just $b - a$. For collections of intervals, the probability is the sum of the probabilities for each interval.

Suppose that the blue region in the spinner goes from 0.05 to 0.2. The probability that the arrow falls into that region is

$$P[0.05 \leq X \leq 0.2] = 0.15$$

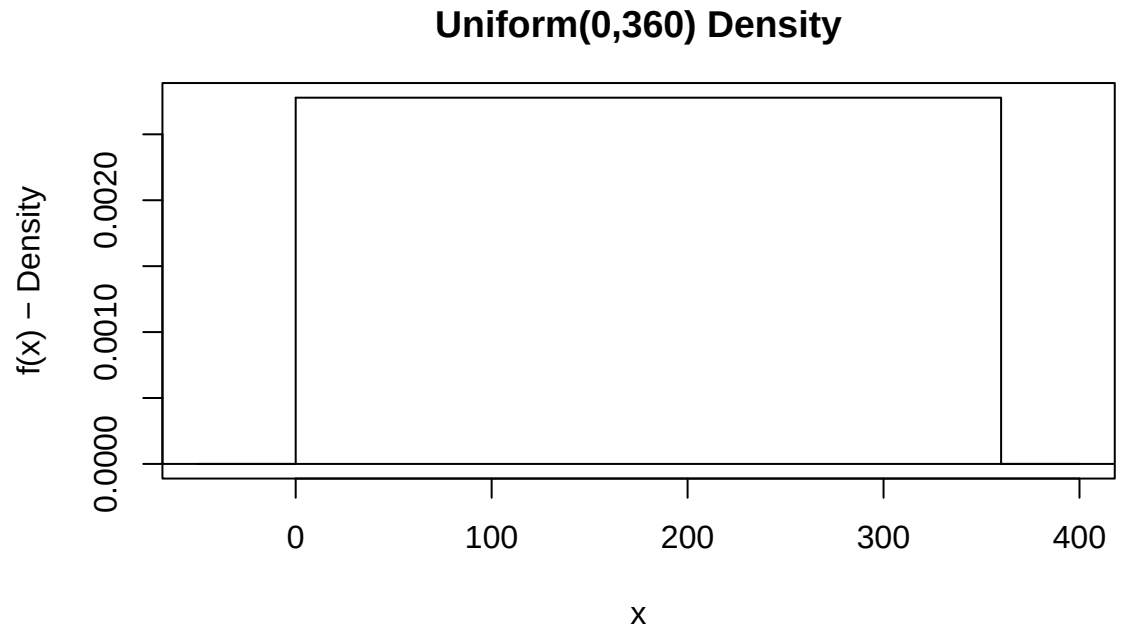




Any other interval of length 0.15 will have the same probability under this model.

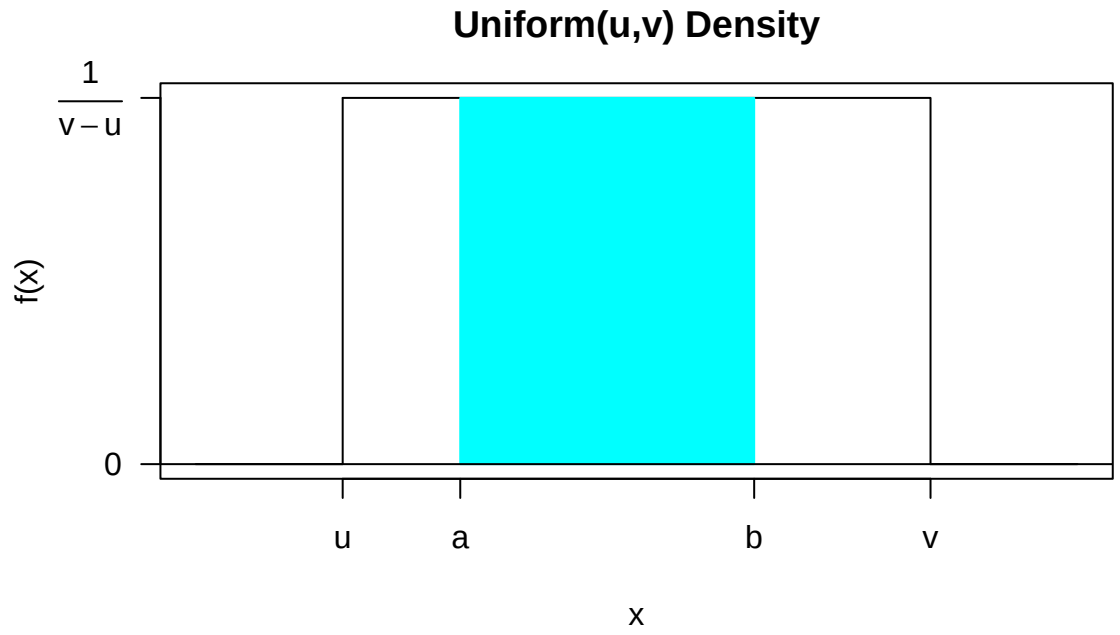
Note that there are many continuous uniform distributions. (One for every possible pair of endpoints for the interval.)

For the spinner example, the position of the arrow could be given as the angle in degrees. The density curve here looks like



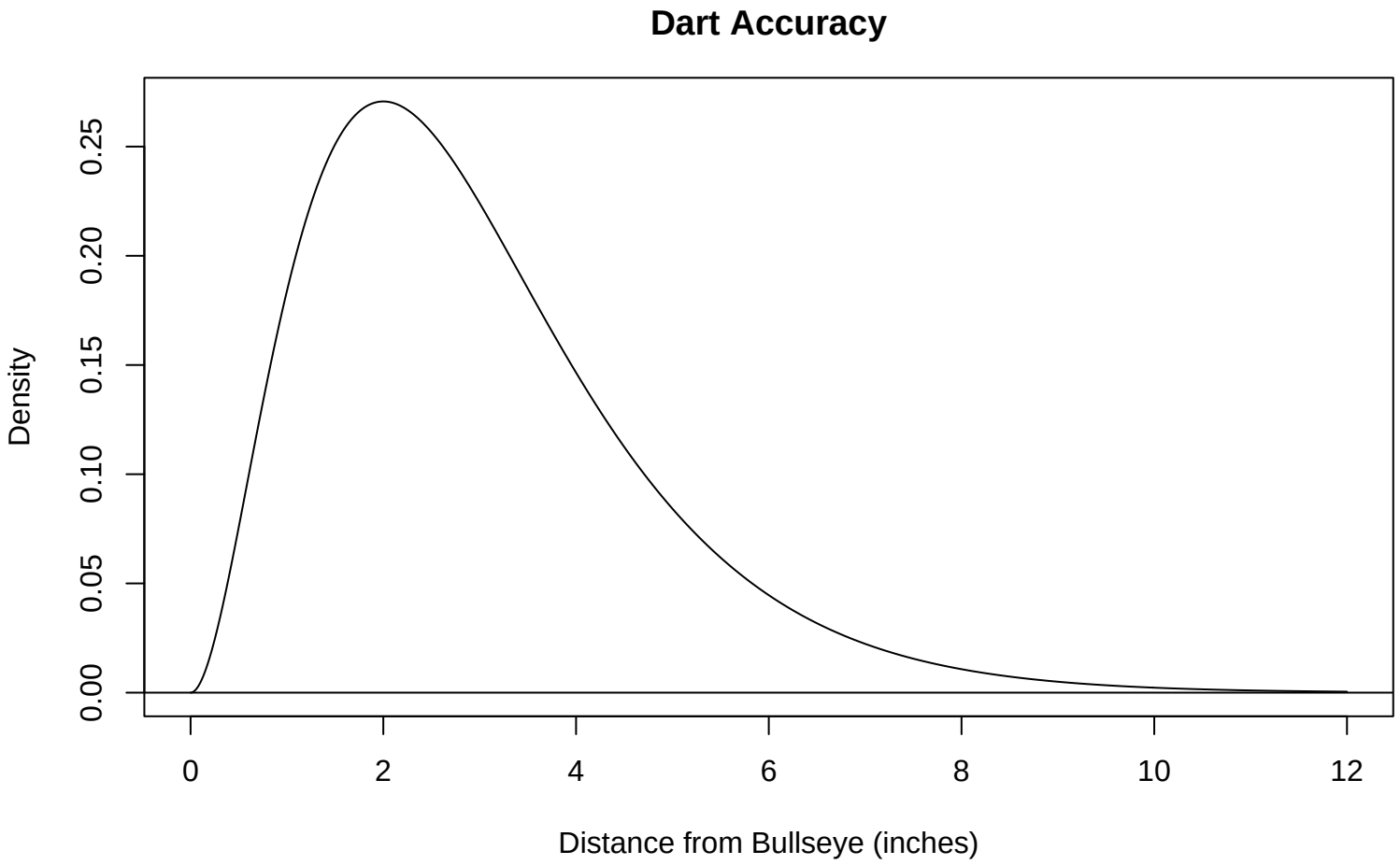
For a uniform distribution with possible values between u and v (sometimes denoted by $Unif(u, v)$), the height of the curve is $\frac{1}{v-u}$ and

$$P[a < X < b] = \frac{b-a}{v-u}$$



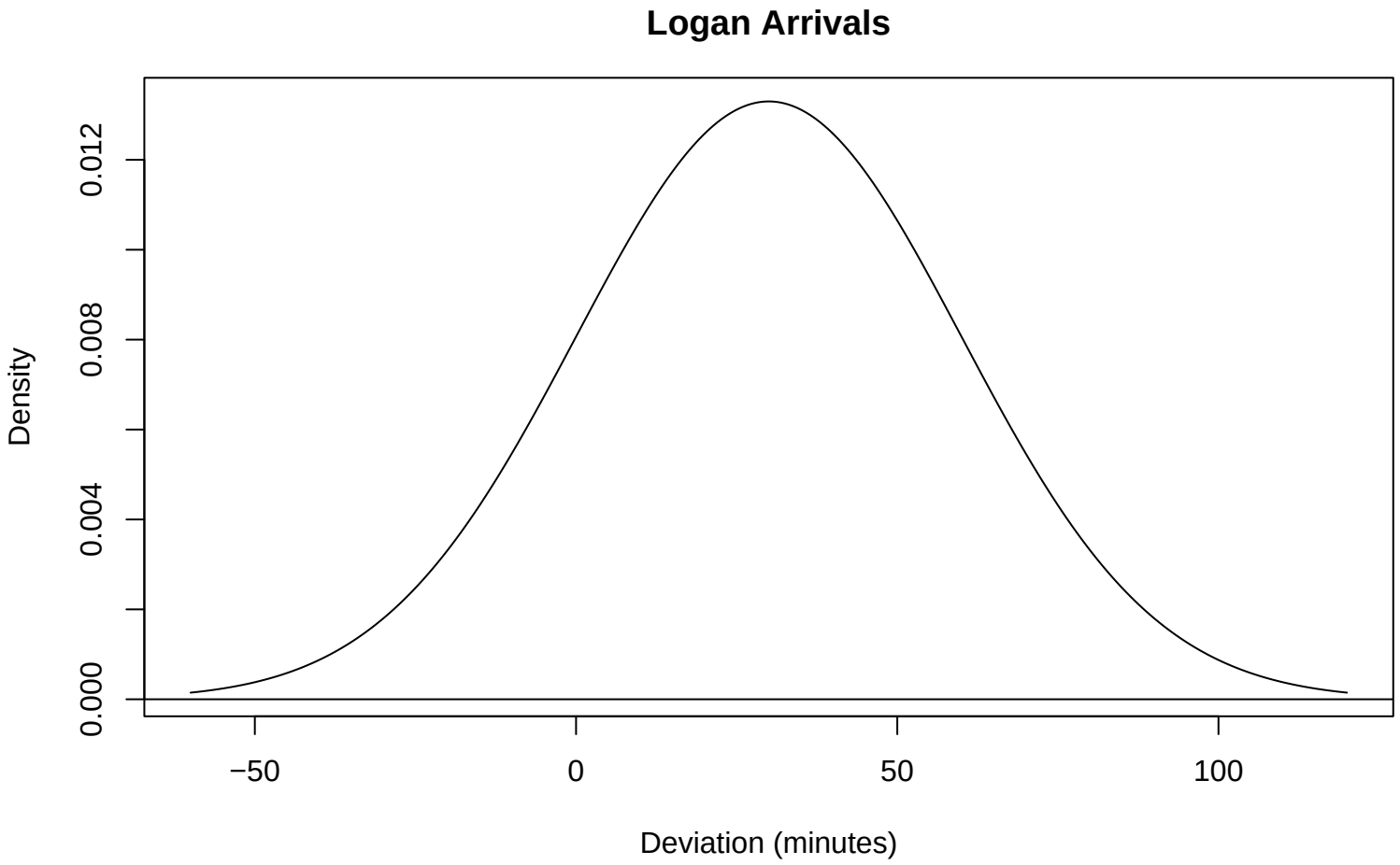
(e.g. height times width)

For the dart board example, the density curve might look like



(The curve is from a Gamma(3,1) distribution)

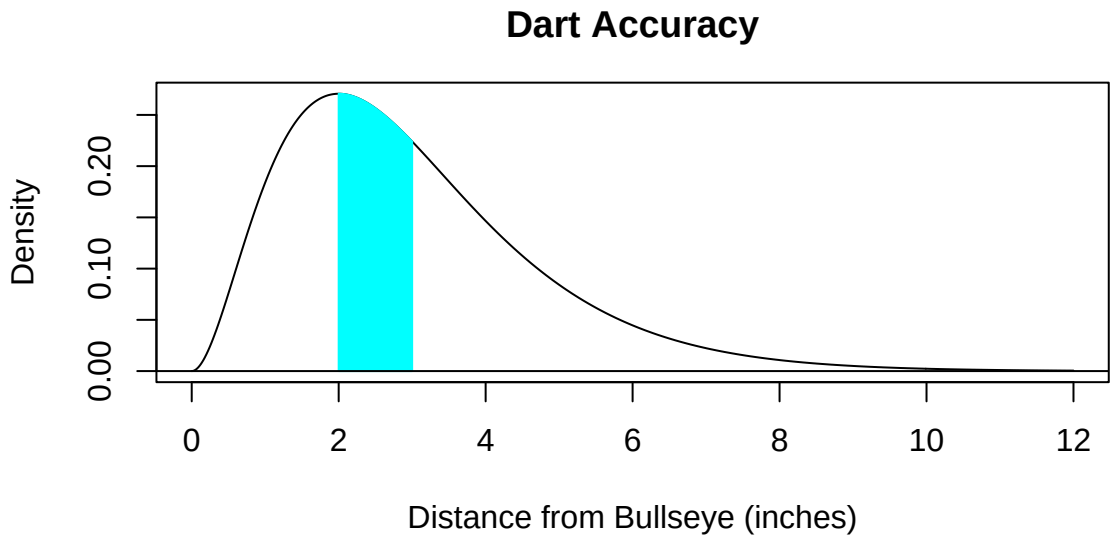
The arrival time deviation at Logan might have a density curve like



The curve is from a Normal(30,30) distribution)

So for the dart example, the probability of a dart being between 2 and 3 inches from the center is the area from 2 to 3 under the curve. Unfortunately, calculating this takes some work. It involves calculus as

$$P[a < X < b] = \int_a^b f(x)dx$$

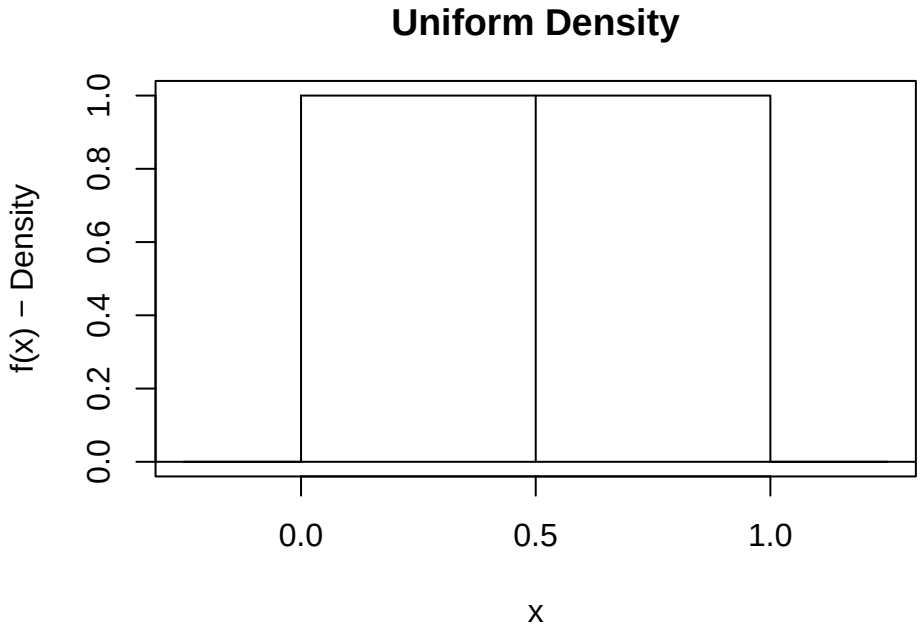


Assuming that the probability model is valid, the probability of a dart begin between 2 and 3 inches from the center is about 0.25.

For most problems, we can use computer programs (as was done for the dart board probability) or tables to calculate probabilities. For many distributions, some special probabilities are tabled which will allow us to determine any probability of the form $P[a < X < b]$. From these, we can get any probability we want.

What is $P[X = a]$?

Let's consider the spinner example and what is $P[X = \frac{1}{2}]$



$$P\left[X = \frac{1}{2}\right] = \text{Area above } \frac{1}{2}$$
$$= 1 \times 0 = 0$$

Any individual value has zero probability (for any continuous RV)

This also implies that for continuous RVs

$$P[X < x] = P[X \leq x]$$
$$P[X > x] = P[X \geq x]$$

However these relationships don't hold for discrete random variables. For example

$$P[X \geq x] = P[X = x] + P[X > x]$$

Cumulative Distribution Function

For calculating probabilities for any distribution, all that is needed is the cumulative distribution function (CDF), particularly for continuous RVs

$$F(x) = P[X \leq x]$$

Inside the front cover of the text is the CDF for the standard normal distribution.

For continuous RVs, there is the following relationship between the CDF and the density curve for a distribution

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(u)du \\ f(x) &= \frac{dF(x)}{dx} \end{aligned}$$

For discrete RVs, the is a similar relationship

$$\begin{aligned} F(x) &= \sum_{i:x_i \leq x} p_i \\ P[X = x] &= P[X \leq x] - P[X < x] \end{aligned}$$