Module C: Understanding the Data-Generating Process

Slide Deck C7:

The Uniform Distribution

The section in which we learn about our first named continuous distribution. Arguably, this is the key distribution from which all other distributions can be named. It describes a continuous random variable in which every element has the same likelihood of happening.

Today's Objectives

By the end of this slidedeck, you should

1. understand the difference between discrete and continuous random variables;
2. know the purpose of the probability *density* function;
3. be able to prove that the density is not a probability;
4. know the purpose of the cumulative distribution function;
5. be able to calculate probabilities using geometry or the CDF; and
6. understand the Uniform distribution in terms of its
   - two parameters,
   - sample space,
   - expected value, and
   - variance.

Continuous RVs

Definition (Continuous Random Variables)

A **continuous random variable** is a random variable with a sample space consisting of an interval of values.

Examples

Examples of continuous random variables:

1. Student height
2. Age of car
3. Time spent at a stop light
4. Distance a golf ball goes

Near Continuous RVs

Frequently, a discrete random variable is "essentially" continuous because the spacing between adjacent values is insignificant when compared to the useful data range.

Examples

Examples of near-continuous random variables:

1. GPA
2. Annual salary
3. Gross domestic product (GDP)
4. GDP per capita
5. Crime rate
6. Number of ears of corn grown in Illinois
7. Number of flu cases in a week

Near Continuous RVs

Why this matters:

- In terms of analysis, there is *no* difference in how continuous and near-continuous variables are handled. Thus, we will ignore the distinction between near-continuous and continuous.
- Statistical procedures dealing with continuous variables are (usually) identical to those dealing with near-continuous variables.
- As such, the near-continuous are treated as continuous for all intents and purposes.

Characteristics of the Uniform

Definition (Uniform Distribution)

The **Uniform distribution** is a continuous distribution that describes random variables whose likelihood of occurring is constant across a specified interval.

If the random variable X has a Uniform distribution, we write:

$$X \sim \mathcal{U}nif(a, b)$$

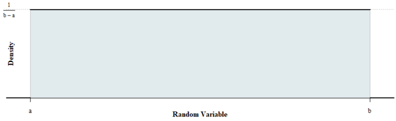
It is characterized by these features

- Two parameters: minimum and maximum values (a and b)
- Sample space is $S = [a, b]$
- $\mathbb{E}\left[X\right] = \frac{a+b}{2}$
- $\mathbb{V}\left[X\right] = \frac{(b-a)^2}{12}$

Probability Density Function

The probability density function for the Uniform distribution is given by

$$f(x;\ a,b) = \begin{cases} \dfrac{1}{b-a} & x \in [a,b] \\ 0 & \text{Otherwise} \end{cases}$$



Purposes of pdf

Note that the probability density function has two purposes:

1. To help the researcher understand probability for a continuous distribution
2. To help the researcher calculate probabilities of a continuous random variable

Important!!

Probabilities are **areas** under the *density* curve.

Example I: Main and Academy

Example

There is only one stop light between home and school in the morning. It regularly cycles among green (175s), yellow (5s), and red (180s). Given that I stop at the light, what is the probability that I wait at most 60 seconds?

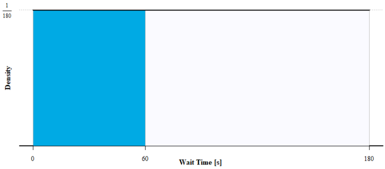
Solution:
Because the time I wait does not depend on when I get there, and because there is a definite lower- and upper-bound to my wait time (0 and 180s), the wait time distribution follows a Uniform distribution. If we define T as the time (in seconds) I spend waiting at this light, then

$$T \sim \mathcal{U}nif(0, 180)$$

We are asked to calculate $\mathbb{P}[\ T \leq 60\ ]$.

Example I: Main and Academy

Graphically, this is what we just described. The large rectangle is the pdf of $\mathcal{U}nif(0, 180)$, and the dark rectangle is the area of $\mathbb{P}[\ T \leq 60\ ]$.



Example I: Main and Academy

Because the pdf of the Uniform distribution is just a rectangle, and because areas in pdfs are probabilities, we just need to calculate the area of the region $T \leq 60$.

This is just a rectangle. Its area is its height times its width.

From geometry:

- Height $= \frac{1}{180}$
- Width $= 60$
- Area (Probability) $= \frac{1}{180} \times 60 = 0.33333\ldots$

Thus, the probability of waiting at most 60s is 33.3%.

Doing this in R just requires the following line:

```
punif(60, min=0, max=180)
```


Cumulative Distribution Function

Definition (Cumulative Distribution Function)

The **cumulative distribution function** (CDF) of a probability distribution is defined as

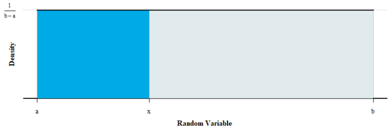
$$F(x) = \mathbb{P}[\ X \leq x\ ]$$

Note that this is the first place we see a probability when deling with continuous random variables. That is because this is the only type of probability that can be calculated with contiuous distributions.

The CDF is frequently difficult to calculate. Usually, it requires integral calculus. However, for the Uniform distribution, we can rely on middle-school geometry.

Uniform CDF

The cumulative distribution function (CDF) is the function that describes $F(x) = \mathbb{P}[\ X \leq x\ ]$. This can be easily calculated for the Uniform distribution, in general.



Remember: Probabilities are just areas under the pdf curve.

Cumulative Distribution Function

The cumulative distribution function (CDF) is the function that describes $F(x) = \mathbb{P}[\ X \leq x\ ]$. This can be easily calculated for the Uniform distribution, in general. Let

$$X \sim \mathcal{U}nif(a, b)$$

The probability $\mathbb{P}[\ X \leq x\ ]$ is the area of the rectangle with

- height $f(x) = \frac{1}{b-a}$
- width: $x - a$

This means the cumulative distribution function is

$$\mathbb{P}[\ X \leq x\ ] = F(x) = \frac{x-a}{b-a}$$

Quantile Function

The CDF starts with a value of the random variable, x, and calculates a cumulative probability, p, such that $\mathbb{P}[\ X \leq x\ ] = p$. The quantile function starts with a cumulative probability, p, and calculates the x-value such that $\mathbb{P}[\ X \leq x\ ] = p$.

The two functions are inverses of each other. That is, $Q(p) = F^{-1}(p)$. Thus, the quantile function for the Uniform distribution is

$$
\begin{aligned}
p &= \frac{x-a}{b-a} \\
p(b-a) &= x-a \\
p(b-a)+a &= x
\end{aligned}
$$

That is, the quantile function for the Uniform distribution is

$$Q(p) = p(b-a)+a$$

Today's Objectives

Now that we have concluded this lecture, you should be able to

1. understand the difference between discrete and continuous random variables;
2. know the purpose of the probability *density* function;
3. be able to prove that the density is not a probability;
4. know the purpose of the cumulative distribution function;
5. be able to calculate probabilities using geometry or the CDF; and
6. understand the Uniform distribution in terms of its
   - two parameters,
   - sample space,
   - expected value, and
   - variance.

Useful R Functions

In this slide deck, we covered these R functions regarding the continuous Uniform distribution:

- `dunif(x, min,max)` is the density function, $f(x)$
- `punif(x, min,max)` is the CDF $F(x) = \mathbb{P}[\ X \leq x\ ]$
- `qunif(p, min,max)` is the quantile function $\mathbb{P}[\ X \leq x\ ] = p$
- `runif(n, min,max)` gives a random sample from this distribution

Please do not forget to use the `allProbabilities` document that lists all of the probability functions in R.

Supplemental Activities

The following are supplements for the topics covered today.

- SCA 6 is for continuous distributions like the Uniform.

Note that you can access all Statistical Computing Activities here:
`https://www.kvasaheim.com/courses/stat200/sca/`

In addition to the SCA, **Laboratory Activity C** is helpful for learning how to handle some continuous distributions (including the Uniform distribution). The lab actually illustrates the Central Limit Theorem.
`https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding continuous distributions, including the Uniform:

- Hawkes Learning: None
- Intro to Modern Statistics: None
- R for Starters: Appendix B.1, B.2
- Wikipedia: Uniform Distribution (Continuous)

Calculus Extra

How does one directly calculate the expected value? Calculus!

In general, the expected value is actually defined as

$$\mathbb{E}\left[\,X\,\right] := \int_S x f(x)\ dx$$

Similarly, the variance is defined as

$$\begin{aligned}\mathbb{V}\left[\,X\,\right] &:= \int_S (x-\mu)^2 f(x)\ dx \\ &= \int_S x^2 f(x)\ dx - \mu^2\end{aligned}$$

Calculus Extra

For the Uniform distribution, the expected value is

$$\begin{aligned}\mathbb{E}\left[X\right] &:= \int_S x f(x)\ dx \\ &= \int_a^b x\frac{1}{b-a}\ dx \\ &= \left[\frac{x^2}{2}\ \frac{1}{b-a}\right|_a^b \\ &= \frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right) \\ &= \frac{1}{2}\ \frac{(b-a)(b+a)}{b-a} \\ &= \frac{a+b}{2}\end{aligned}$$

Calculus Extra

For the Uniform distribution, the variance is

$$\begin{aligned}\mathbb{V}\left[X\right] &= \int_a^b x^2\frac{1}{b-a}\ dx - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{b-a}\ \left[\frac{x^3}{3}\right|_a^b - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{3}\frac{b^3-a^3}{b-a} - \frac{1}{4}\left(a^2+b^2+2ab\right) \\ &= \frac{1}{12}\left(\frac{4(b-a)(b^2+ab+a^2)}{b-a} - 3\left(a^2+b^2+2ab\right)\right) \\ &= \frac{1}{12}\left(4b^2+4ab+4a^2-3a^2-3b^2-6ab\right) \\ &= \frac{1}{12}\left(b^2-2ab+a^2\right) \qquad = \frac{(b-a)^2}{12}\end{aligned}$$