Chapter 10

The Uniform Distribution

References: Pruim 3.1.1, Larsen & Marx 3.4

The uniform distribution is the continuous equivalent of the equiprobable probability model on a discrete sample space.

Definition 10.1. A continuous **uniform** random variable on the interval $[a, b)$ is the random variable with PDF

$$f(x) = \frac{1}{b-a}, \quad a \leq x < b.$$

We will denote this by $X \sim Unif(a, b)$.

It's easy to verify that f is a valid PDF either with geometry or using integration.

. . . .

Example 10.1. Let $X \sim Unif(0, 10)$.

a. Find $P(3 \leq X < 7)$?

b. The median - or 50th percentile - of a continuous random variable is the number m such that

$$P(X \leq m) = \frac{1}{2}.$$

Find the median of X.

. . . .

The R functions `dunif` and `punif` work on the same principles we have been using for discrete distributions.

```
dunif(x = 3, min = 0, max = 10) # f(x)
## [1] 0.1

punif(q = 7, min = 0, max = 10, lower.tail = F) # P(X > 7)
## [1] 0.3

diff(punif(q = c(3, 7), min = 0, max = 10)) # P(3 <= X < 7)
## [1] 0.4
```

In addition, the function `qunif` can be used to find percentiles of a uniform distribution. That is, the function `qunif` finds x such that

$$F(x) = p.$$

```
qunif(p = 0.5, min = 0, max = 10 )    #solve F(x) = 0.5
```

```
## [1] 5
```

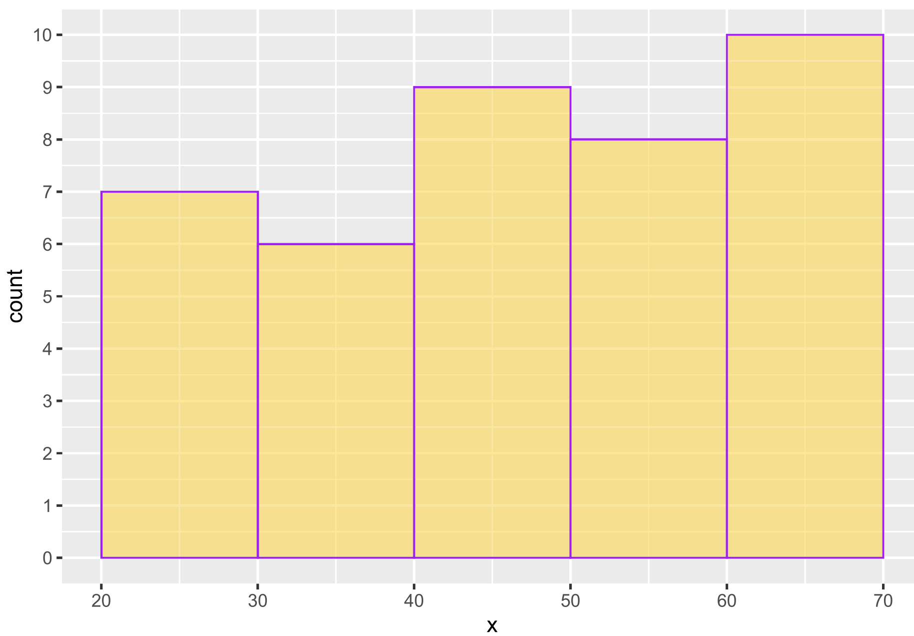
10.1 Fitting a PDF to data

Let's now consider the problem of using a continuous probability function to model the distribution of a set of n measurements: $x_1, x_2, \ldots, x_n$. Following the approach taken with **Fumbles** data, we would start by making a histogram of the n observations. The problem, though, is that the sum of the areas of the bars comprising that histogram would not necessarily equal 1.

As a case in point the following shows a set of forty observations. Grouping those xi 's into five classes, each of width 10, produces the distribution and histogram pictured below.

```
sample_data <- data.frame(
  x = c( 33.8, 62.6, 42.3, 62.9, 32.9, 58.9, 60.8, 49.1,
         42.6, 59.8, 41.6, 54.5, 40.5, 30.3, 22.4, 25.0,
         59.2, 67.5, 64.1, 59.3, 24.9, 22.3, 69.7, 41.2,
         64.5, 33.4, 39.0, 53.1, 21.6, 46.0, 28.1, 68.7,
         27.6, 57.6, 54.8, 48.9, 68.4, 38.4, 69.0, 46.6 ) )

ggplot(data = sample_data, mapping = aes( x =x ) ) +
  geom_histogram(binwidth = 10,     #Sturges rule: log_2(n)+1
                 breaks=seq(20,70,10),
                 alpha = 0.5,
                 color = "purple",
                 fill = "gold" )+
  scale_y_continuous(breaks=0:10)
```



Furthermore, suppose we have reason to believe that these forty xi 's may be a random sample from a uniform probability function defined over the interval $[20, 70)$ - that is,

$$f(x) = \frac{1}{50} \quad 20 \leq x < 70.$$

How can we appropriately draw the distribution of the xi 's and the uniform probability model on the same graph to assess the fit?

Note, first, that $f(x)$ and the histogram are not compatible in the sense that the area under $f(x)$ is (necessarily) 1, but the sum of the areas of the bars making up the histogram is 400:

$$7 \times 10 + 6 \times 10 + 9 \times 10 + 8 \times 10 + 10 \times 10$$

Nevertheless, we can "force" the total area of the five bars to match the area under $f(x)$ by redefining the scale of the vertical axis on the histogram. Specifically, frequency needs to be replaced with the analog of probability density, which would be the scale used on the vertical axis of any graph of $f(x)$.

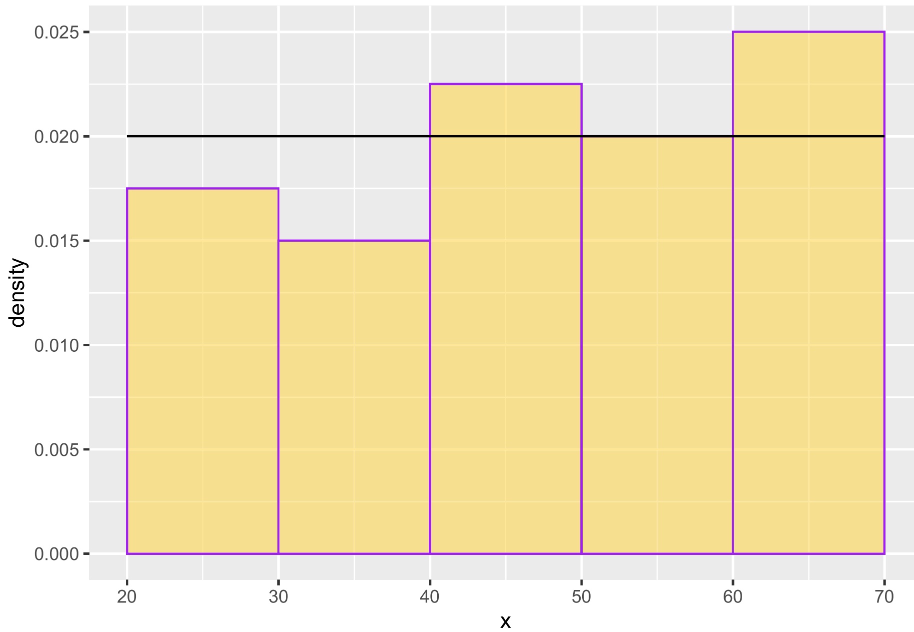
The following figure shows the histogram of the data where the height of each bar has been converted to the density scale according to the formula:

$$\text{density (of a class)} = \frac{\text{class frequency}}{n \times \text{class width}}$$

where n is the number of observations in the dataset: $n = 40$.

Superimposed is the PDF of the uniform distribution $f(x) = \frac{1}{50} \quad 20 \leq x < 70$. We can use the familiar `geom_function` layer to add a theoretical probability distribution as shown below.

```
ggplot( ) +
  geom_histogram(data = sample_data,
                 mapping = aes(x = x,
                               y = after_stat(density)),
                 binwidth = 10,
                 breaks=seq(20,70,10),
                 alpha = 0.5,
                 color = "purple",
                 fill = "gold" )+
  geom_function(fun = dunif,
                args = list(min = 20, max = 70 ),
                xlim = c(20,70) )
```



In practice, density-scaled histograms offer a simple, but effective, format for examining the "fit" between a set of data and a presumed continuous model.

10.2 Practice Problems

1. Suppose $X \sim Unif(-1, 1)$. Find $P(|X| > 1/2)$.
2. You arrive at a bus stop at 10 o'clock knowing that the bus will arrive at some time uniformly distributed between 10 and 10:30.

a. What is the probability that you will wait longer than 10 minutes?

b. If, at 10:15 the bus has not yet arrived, what is the probability that you will have to wait at least an additional 10 minutes?