Chapter 10

Sunday, November 12, 2023 11:55 AM

Chapter 10

The Uniform Distribution

The uniform experiment

Select a number randomly from the interval $[a, b)$.

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

Probability Density Function

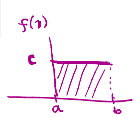
A continuous **uniform** random variable on the interval $[a, b)$ is the random variable with PDF

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

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

Since every number in $[a, b)$ is equally probable ($\because$ randomly selecting), it makes sense that

$$f(x) = c \qquad a \leqslant x < b$$



In order for $f(x)$ to be a valid PDF, we need $c > 0$ and

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

$$\Rightarrow \int_a^b c\, dx = c\, x \Big|_a^b = c(b-a) = 1$$

$$\Rightarrow c = \frac{1}{(b-a)}$$

Example 10.1

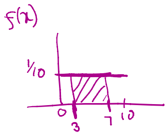
$a \geqslant\ \leftarrow b$

Example 10.1

Let $X \sim Unif(0, 10)$. a b

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

$f(x) = \frac{1}{10}$ $\quad 0 \leqslant x < 10$



$P(3 \leqslant X < 7) = \int_3^7 f(x)\, dx$

$= \int_3^7 \frac{1}{10}\, dx = \frac{1}{10} \cdot x \Big|_3^7 = \frac{1}{10}(7-3) = \boxed{\frac{4}{10}}$

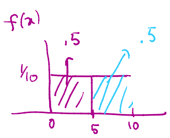
Example 10.1

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

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

$$P(X \leqslant q) = P(X \geqslant q) = \frac{1}{2}.$$

Find the median of X.



$P(X \leqslant q) = \int_{-\infty}^{q} f(x)\, dx = \int_0^q \frac{1}{10}\, dx$

$= \frac{1}{10} \cdot x \Big|_0^q = \frac{q}{10}$

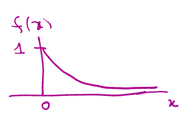
We want q such that $\frac{q}{10} = \frac{1}{2} \Rightarrow q = 5$

$X \sim Binom(n=3, \frac{1}{2})$

x	0	1	2	3
$f(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Variation: Find the median for the PDF

$f(x) = e^{-x} \quad 0 \leqslant x < \infty$



The median q satisfies

$P(X < q) = \frac{1}{2} = P(X > q)$

$P(X < q) = \int_{-\infty}^{q} f(x)\, dx = \int_0^q e^{-x}\, dx = -e^{-x}\Big|_0^q$

$= -e^{-q} - (-1) = 1 - e^{-q}$

Solve $1 - e^{-q} = \frac{1}{2}$ for q. $\Rightarrow e^{-q} = \frac{1}{2} \Rightarrow \ln e^{-q} = \ln\frac{1}{2} \Rightarrow -q = \ln\frac{1}{2}$

$\Rightarrow q = -(\ln \frac{1}{2})$

Uniform PDF calculations in R

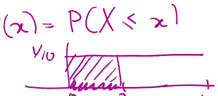
f(3)

$f(x) = \frac{1}{10} \quad 0 \leqslant x < 10$

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

punif(q = 3, min = 0, max = 10)      #P(X <= 3), P(X < 3)
## [1] 0.3
```

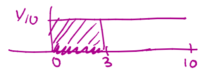
$F(x) = P(X \leqslant x)$

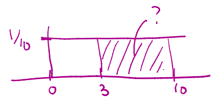


```r
punif(q = 3, min = 0, max = 10)      #P(X <= 3), P(X < 3)
## [1] 0.3

punif(q = 7, min = 0, max = 10)      #P(X < = 7), P(X < 7)
## [1] 0.7

qunif(p = 0.5, min = 0, max = 10 )   #pth percentile
## [1] 5
```





> 1 - punif(q=3, min=0, max=10)

> punif(q=3, min=0, max=10, lower.tail=F)

$P(3 \leq X < 7) = 0.7 - 0.3$

$= \text{diff}(\text{punif}(q = c(3,7), \text{min}=0, \text{max}=10))$

Fitting a PDF to data

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,

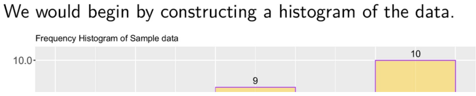
$$f(x) = \frac{1}{50} \quad 20 \leq x < 70.$$

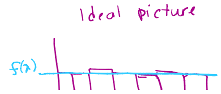
```r
sample_data <- tibble(
    xval  = c( 33.8, 62.6, 42.3, 62.9, 32.9, 58.9, 60.8, 49.1,
          42.6, 59.8, 41.6, 54.5, 40.5, 30.3, 22.4, 25.0,
          59.2, 67.5, 64.1, 59.3, 24.9, 22.3, 69.7, 41.2,
          64.5, 33.4, 39.0, 53.1, 21.6, 46.0, 28.1, 68.7,
          27.6, 57.6, 54.8, 48.9, 68.4,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

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

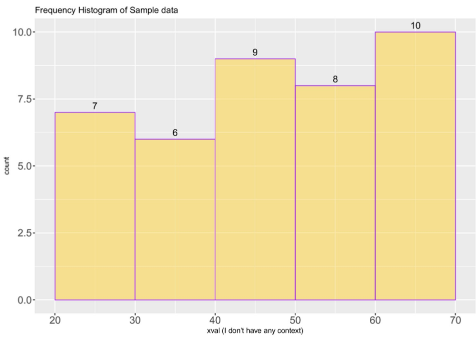
We would begin by constructing a histogram of the data.

Ideal picture

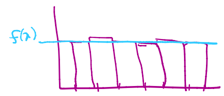




We would begin by constructing a histogram of the data.



Ideal picture



Note, first, that the uniform PDF $f(x)$ 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:

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

Rescaling: 40x10 40x10 40x10 40x10 40x10

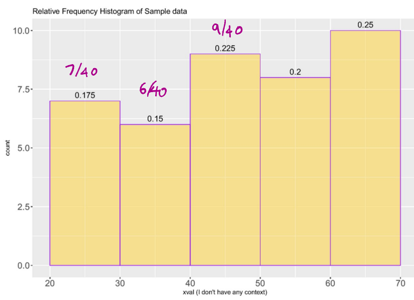
n (# obs) bin width

If we divide each count by n x binwidth then area of the bars = 1.

A first idea to make them compatible is to re-scale the y axis to be a relative frequency:

$$\text{relative freq } = \frac{\text{freq}}{n}$$

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



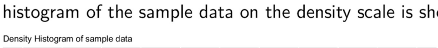
The histogram of the data is still not compatible with a uniform PDF because the total area of the blocks is still not 1.

To make them compatible, we need to use

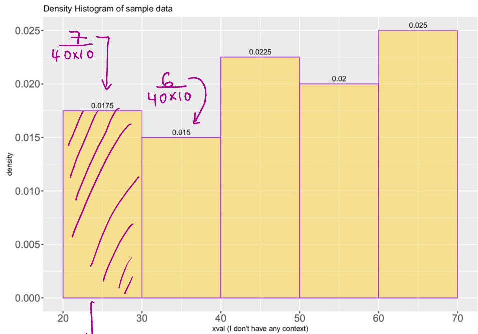
$$\text{density (of a bin)} = \frac{\text{relative frequency}}{\text{width}}$$

on the y axis.

The histogram of the sample data on the density scale is shown below.



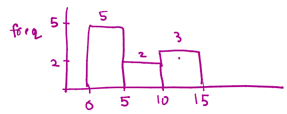
The histogram of the sample data on the density scale is shown below.



area $= \frac{7}{40 \times 10} \times 10 = \frac{7}{40}$

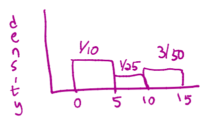
In a density histogram, the area of each block is the relative frequency of the block.

Another example:



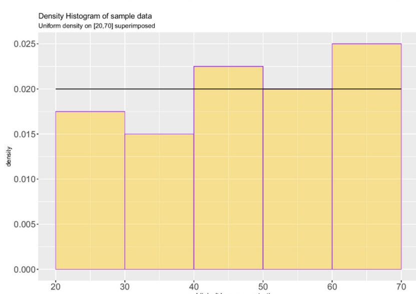
$\boxed{n = 10}$ obs.

area of blocks:
$5 \times 5 + 2 \times 5 + 3 \times 5 = \boxed{50}$



area of blocks:
$\frac{1}{10} \times 5 + \frac{1}{25} \times 5 + \frac{3}{50} \times 5$
$= 1$

And here is the density histogram of the data with the uniform PDF overlaid. Just eyeballing, is the uniform model a good fit?



Code for making the density histogram of the data with uniform density overlaid

```
## Binwidth: A rough rule of thumb for picking the binwidth is to use the range (max - min) of the data
## divided by log_2(n) + 1 where n is number of obs.


ggplot() +
  geom_histogram(data = sample_data,
                 mapping = aes(x = xval,
                               y = after_stat(density)),
                 breaks=seq(20,70,10), → binwidth= 10
                 alpha = 0.5,
                 color = "purple",
                 fill = "gold" )+
  geom_function(fun = dunif,
                args = list(min = 20, max = 70 ),
                xlim = c(20,70) ) +
  labs(x = "xval (I don't have any context)",
       title = "Density Histogram of sample data ",
       subtitle = "Uniform density on [20,70] superimposed")+
  theme(axis.text=element_text(size=15))
```