Econ 5023: Statistics for Decision Making

Univariate Statistics (III): Continuous Variables

Le Wang

Question:

How should we obtain the best forecast for a continuous variabe?

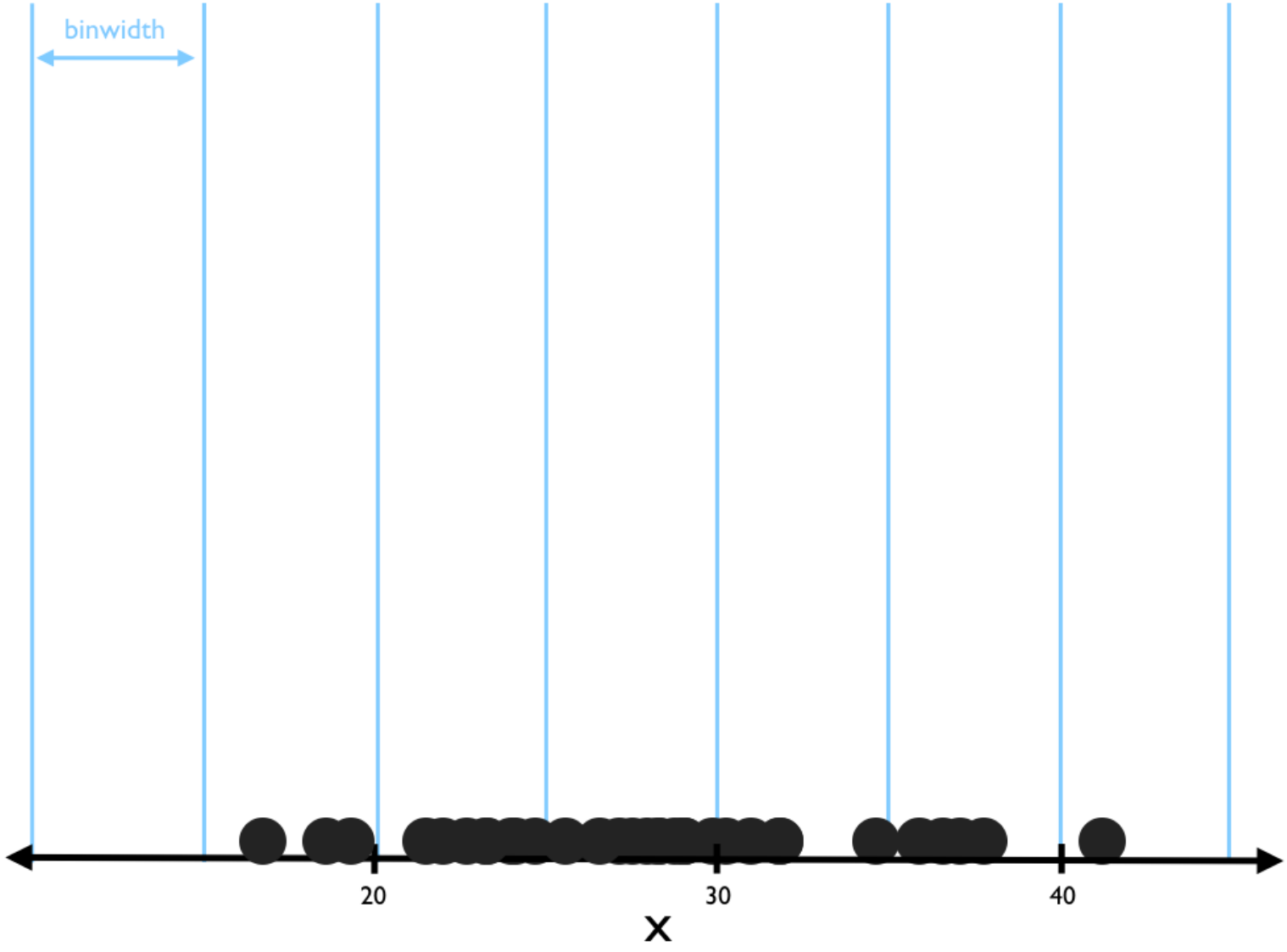
Issues:

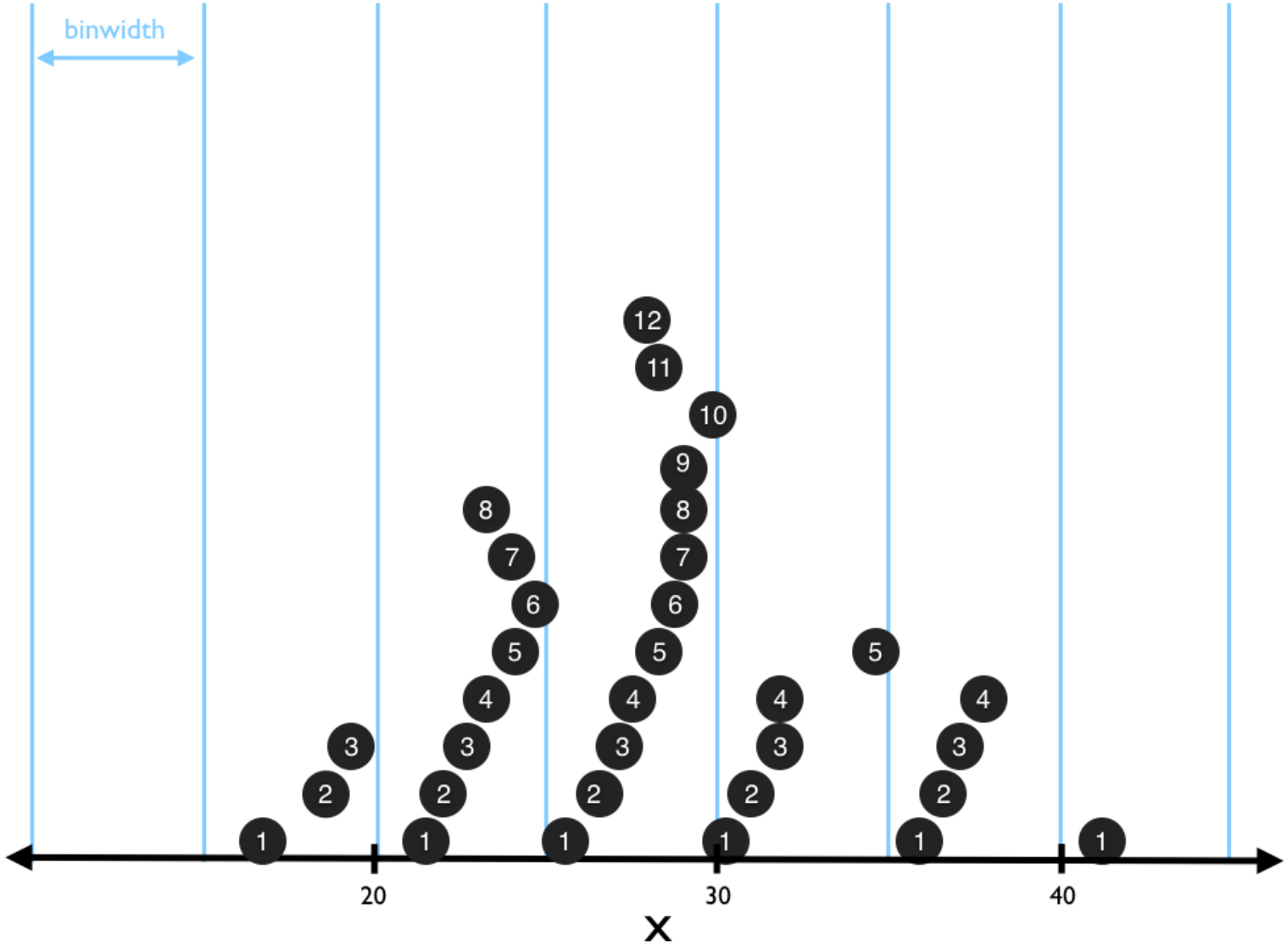
Conceptually we do not know the probability for a particular value. Why?

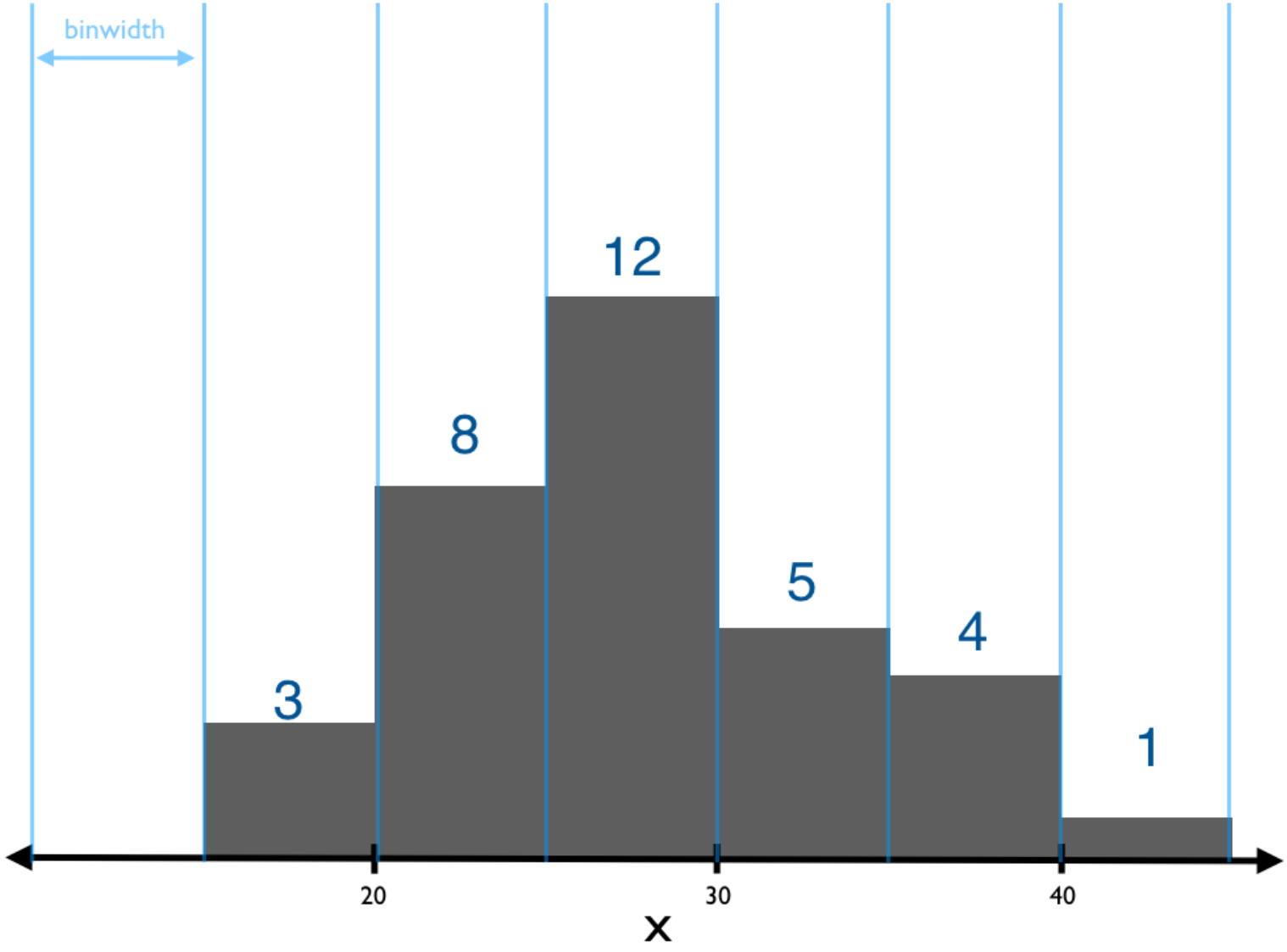
Issues:

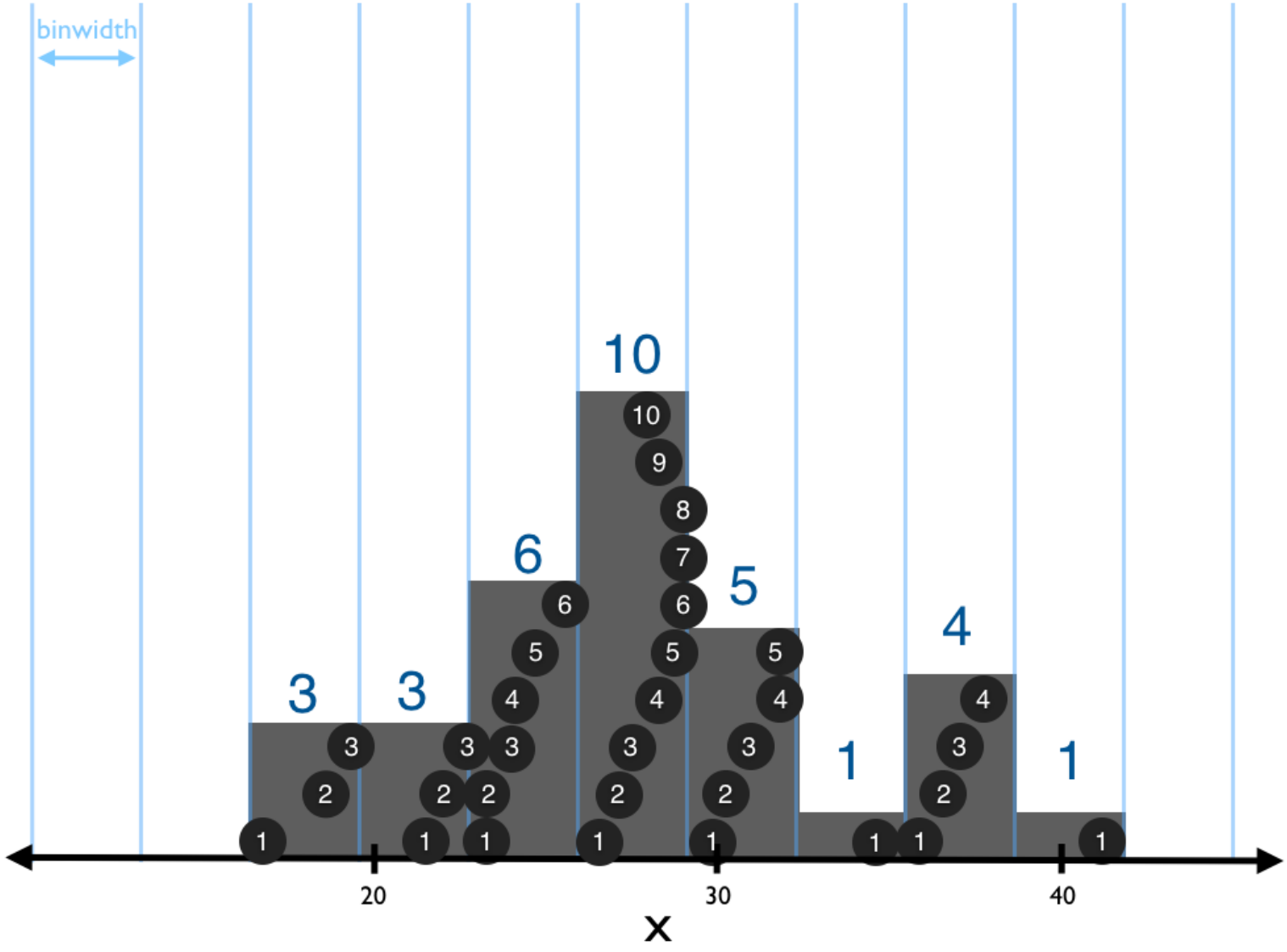
Conceptually we do not know the probability for a particular value.
Why?
A continuous variable can take on an infinite number of values!

Approach 1 (Complete Distribution Approach but approximation): Discretize the continuous variable


binwidth
20
30
40
X


binwidth
12
11
10
9
8
8
7
7
6
6
5
5
5
4
4
4
4
3
3
3
3
2
2
2
2
1
1
1
1
1
20
30
40
X


binwidth
12
8
5
4
3
1
20
30
40
x


binwidth
3
3
6
10
5
1
4
1
20
30
40
x

Approach 1 (Complete Distribution Approach but approximation):
Discretize the continuous variable
Plot the distribution for a continuous variable:

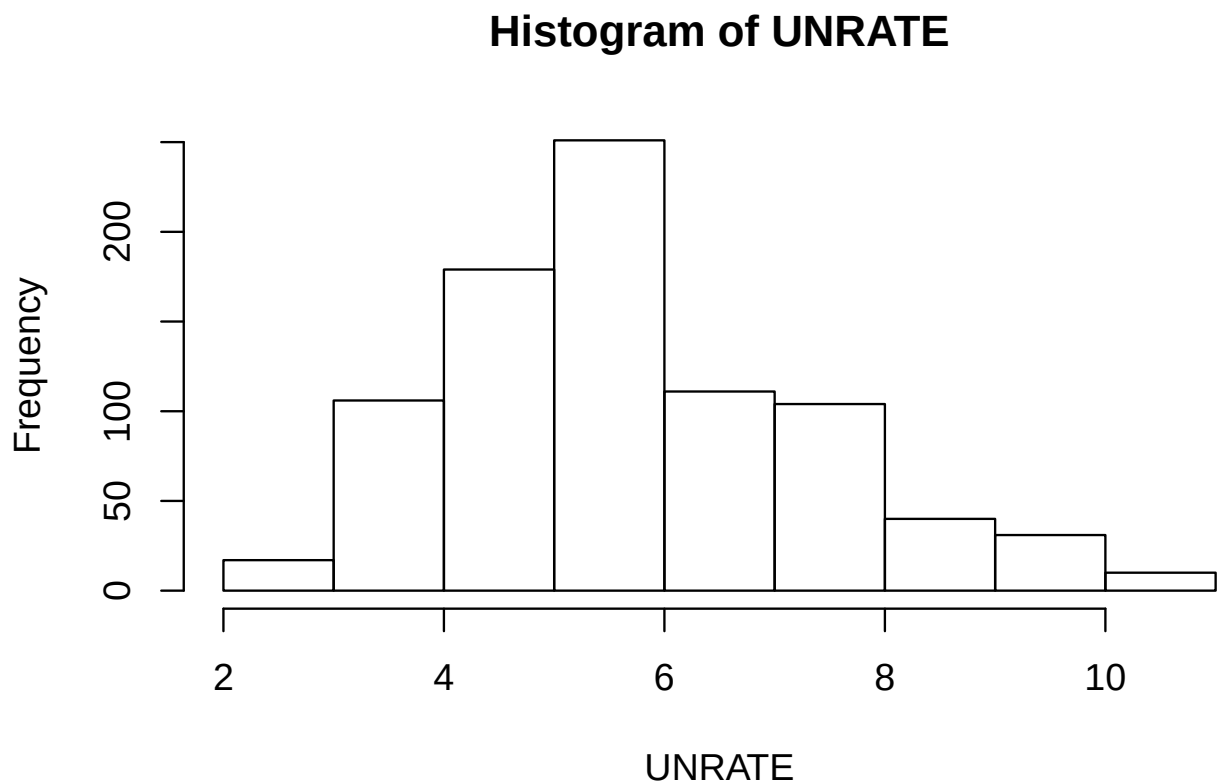
Frequency Distribution: A grouping of quantitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

Nothing but treating a continuous variable as a discrete one!

```
  library(quantmod)
  getSymbols('UNRATE',src='FRED')
```

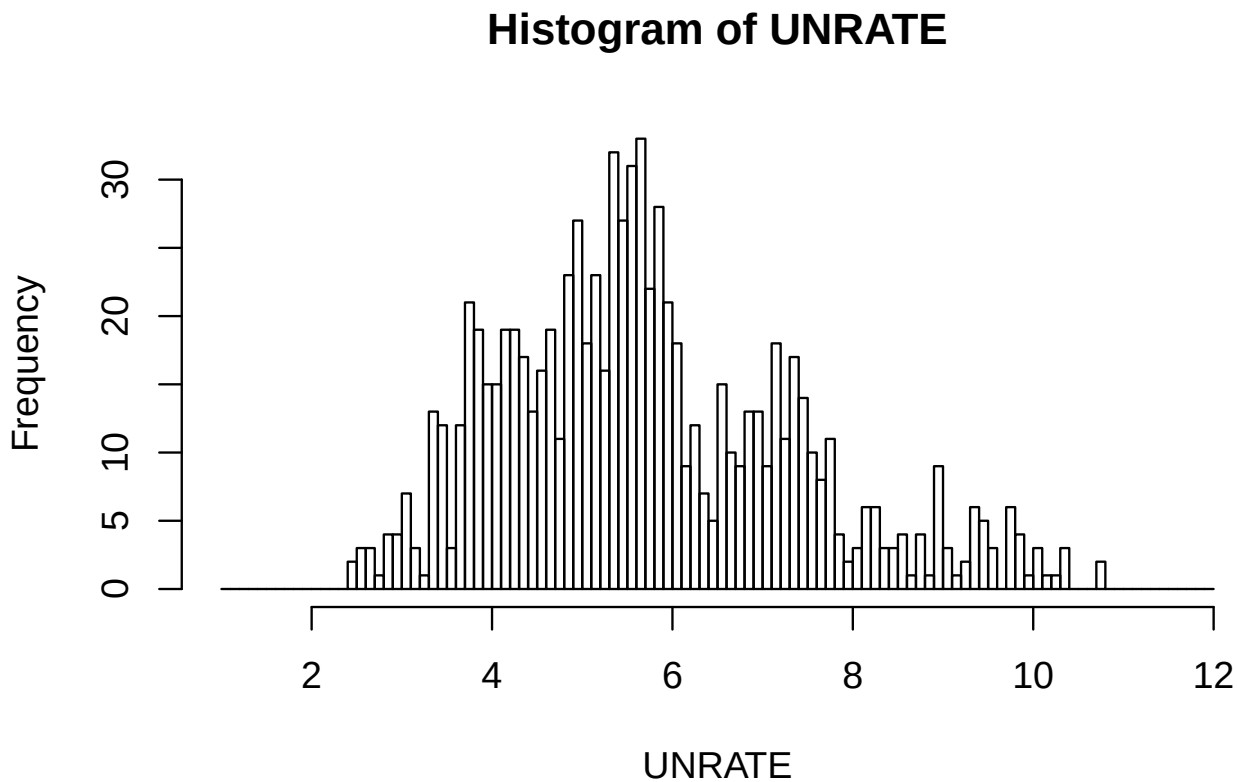
```
## [1] "UNRATE"
```

```
  hist(UNRATE)
```

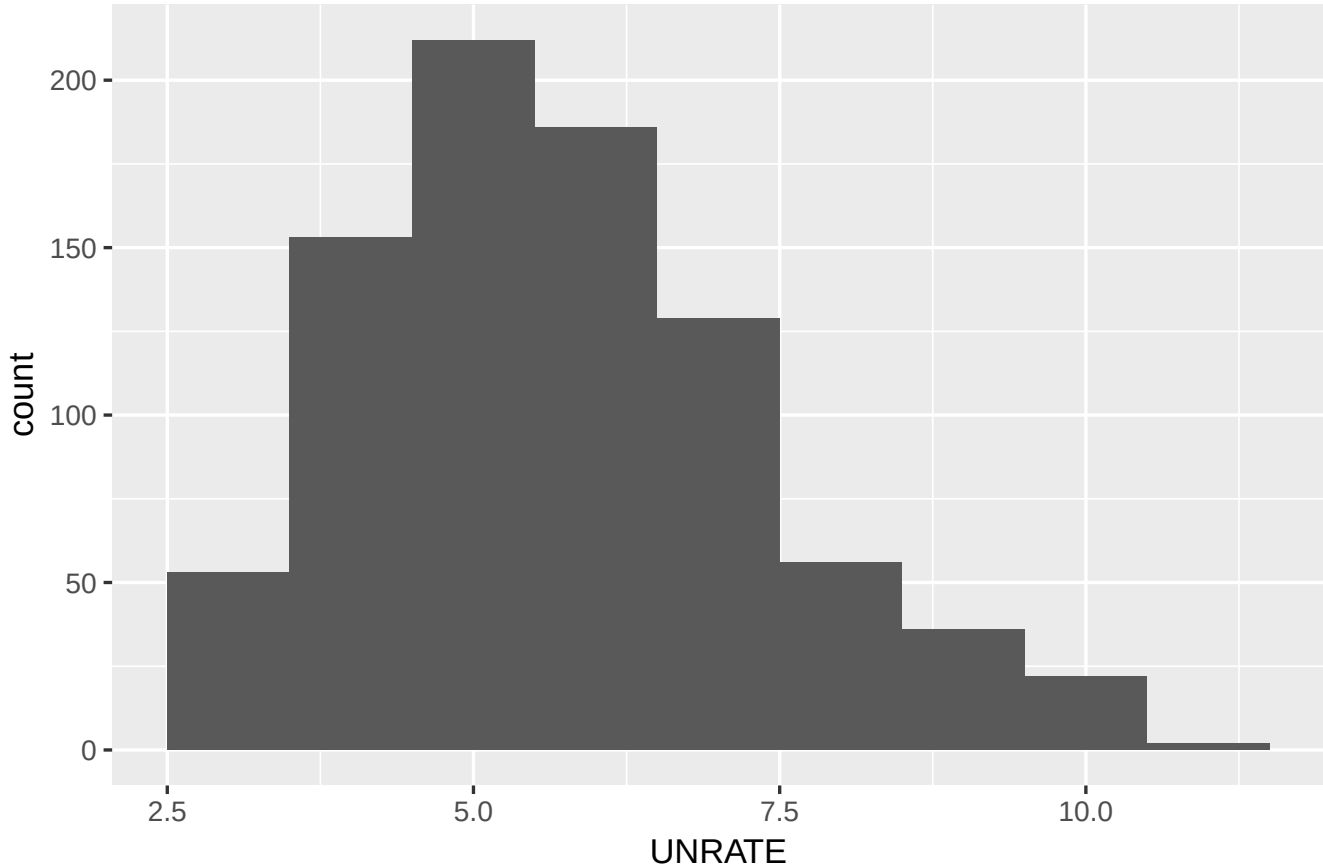


You can even refine it further:

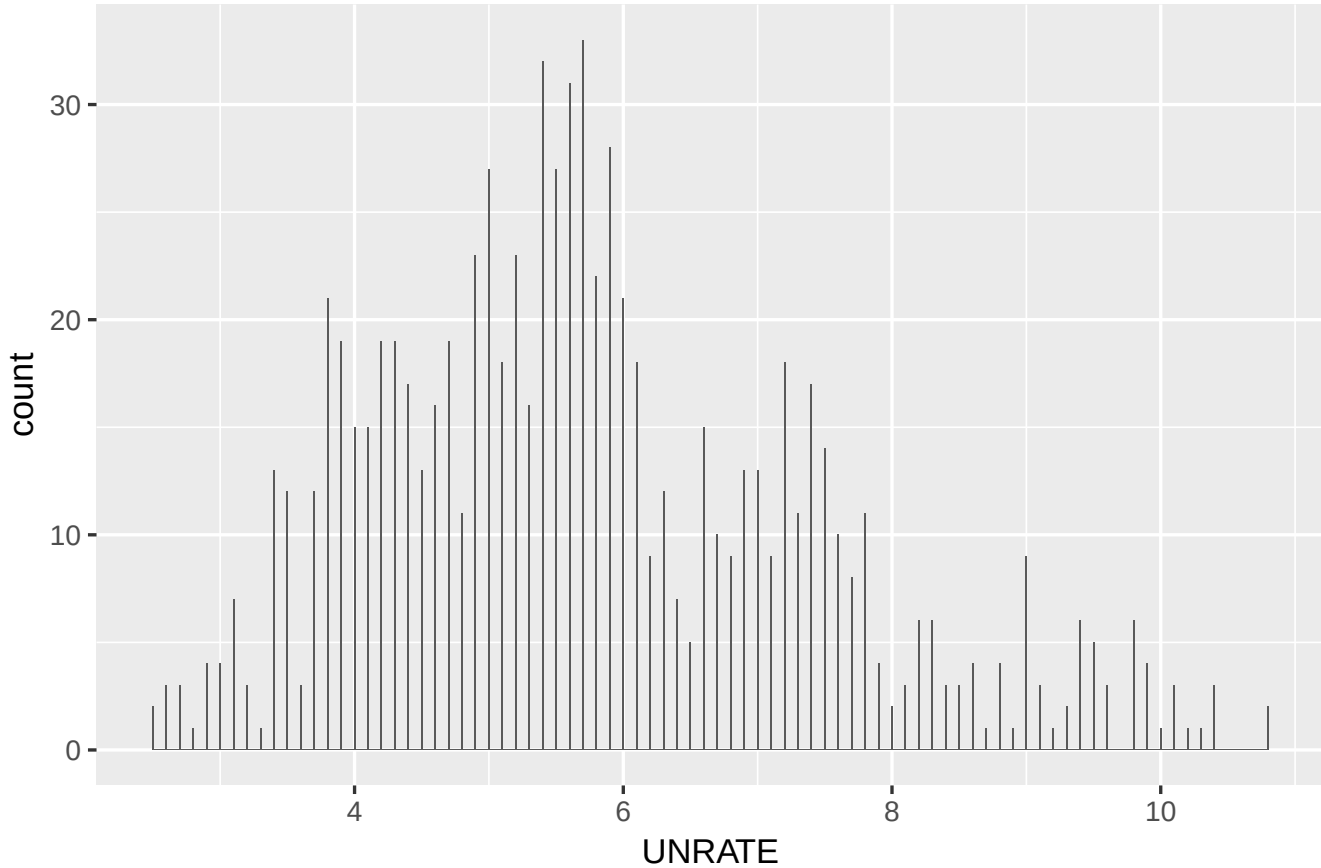
```
hist(UNRATE, breaks = seq(1,12,.1), freq = TRUE)
```



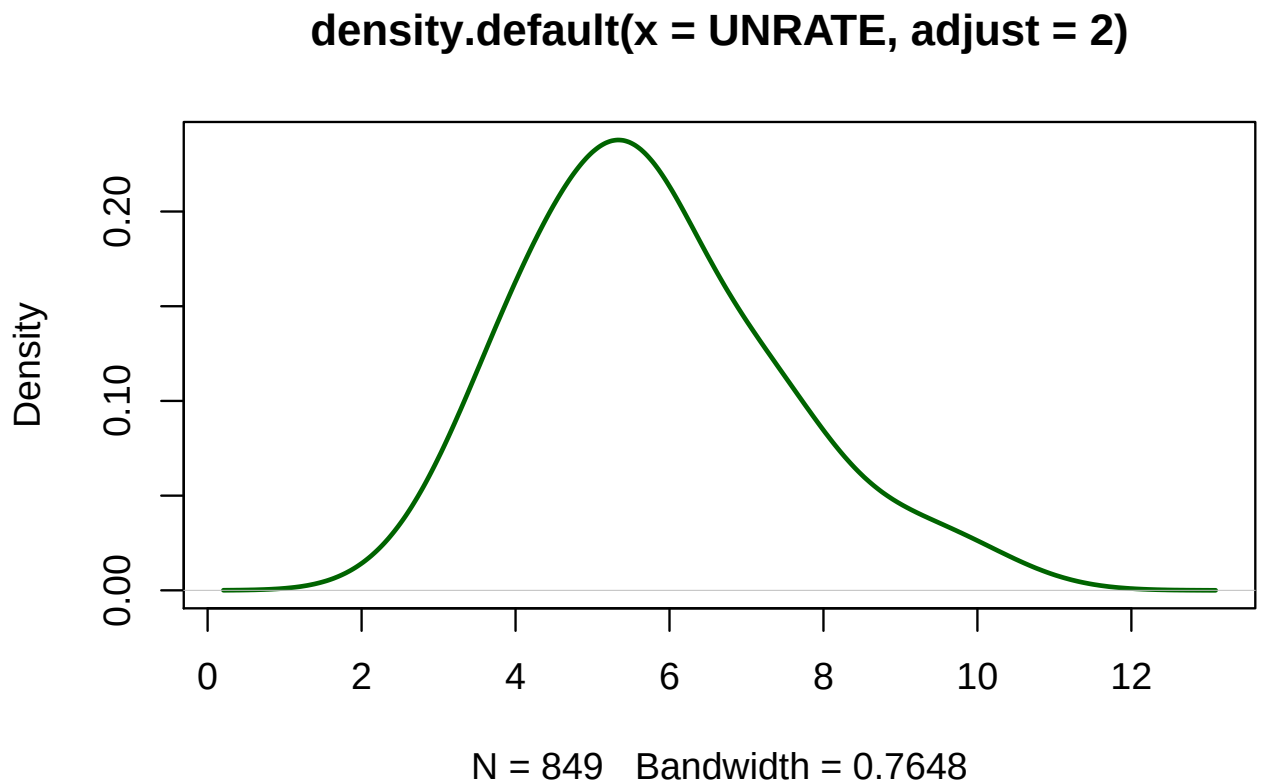
```
library(ggplot2)
ggplot(data = UNRATE, aes(x=UNRATE)) +
  geom_histogram(binwidth = 1)
```



```
library(ggplot2)
ggplot(data = UNRATE, aes(x=UNRATE)) +
  geom_histogram(binwidth = .01)
```



You can even refine it further: **Smoothed Version: Probability Density Function**



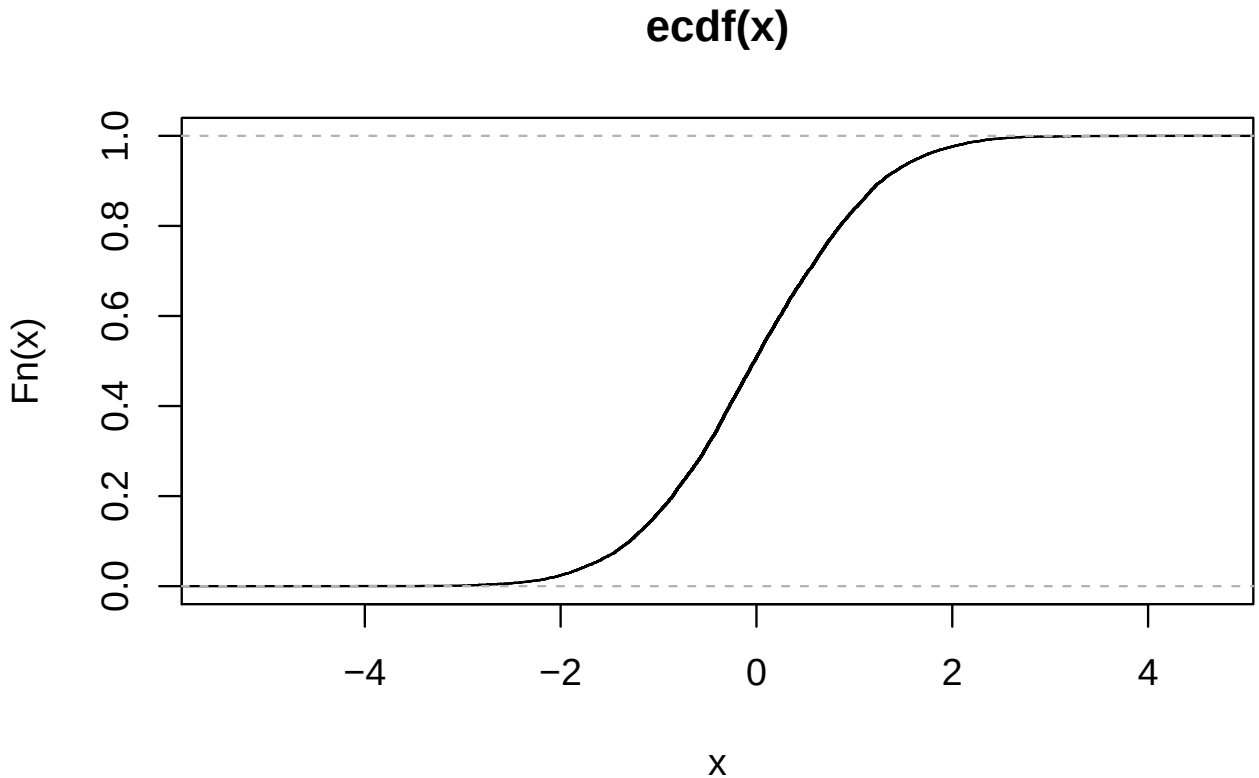
The probability density function, $f(x)$, provides a **relative probability** for a particular value.

CDF: $F(x) = \Pr[X \leq x] = \int_{-\infty}^{x} f(t)dt$

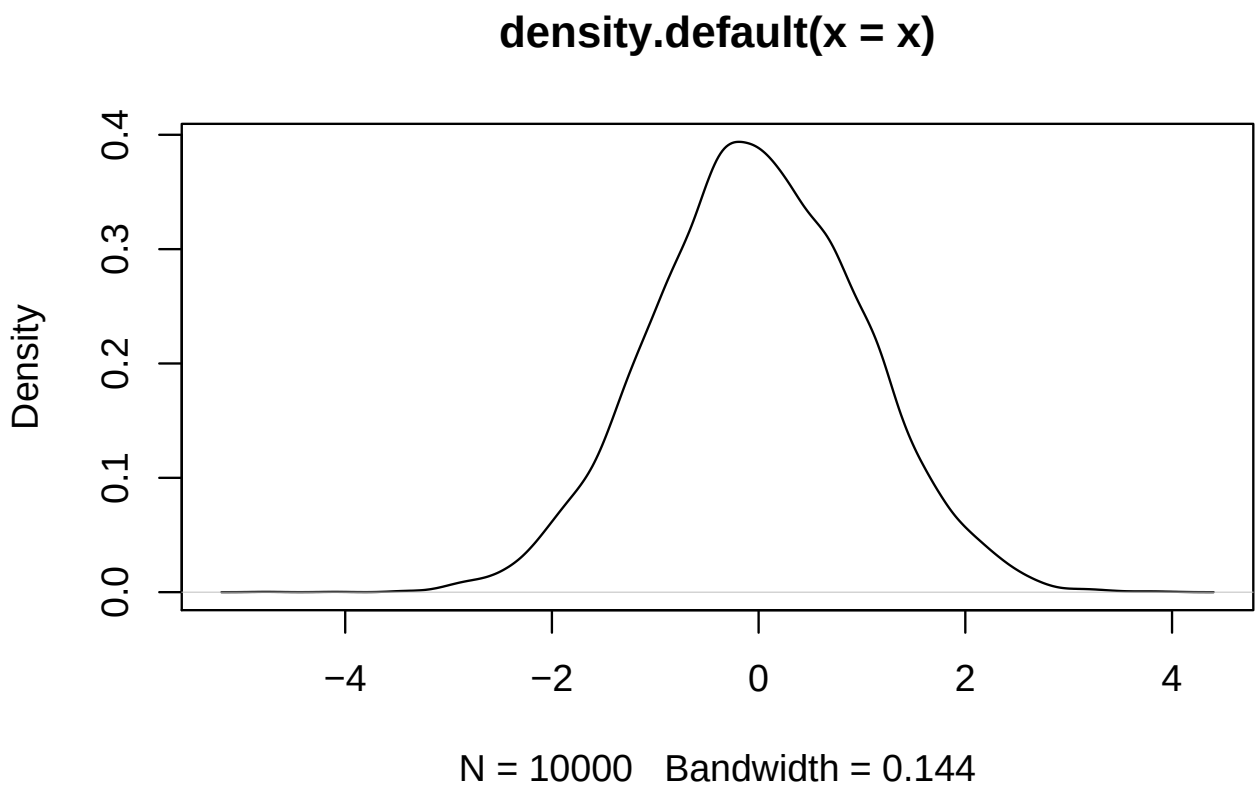
PDF (if contionuous at x): $f(x) = \frac{d}{dx}F(x)$

Remember: in the case of discrete variables, the PMF can be obtained as the changes in CDF with respect to changes in X.

```
plot(ecdf(x))
```



```
plot(density(x))
```



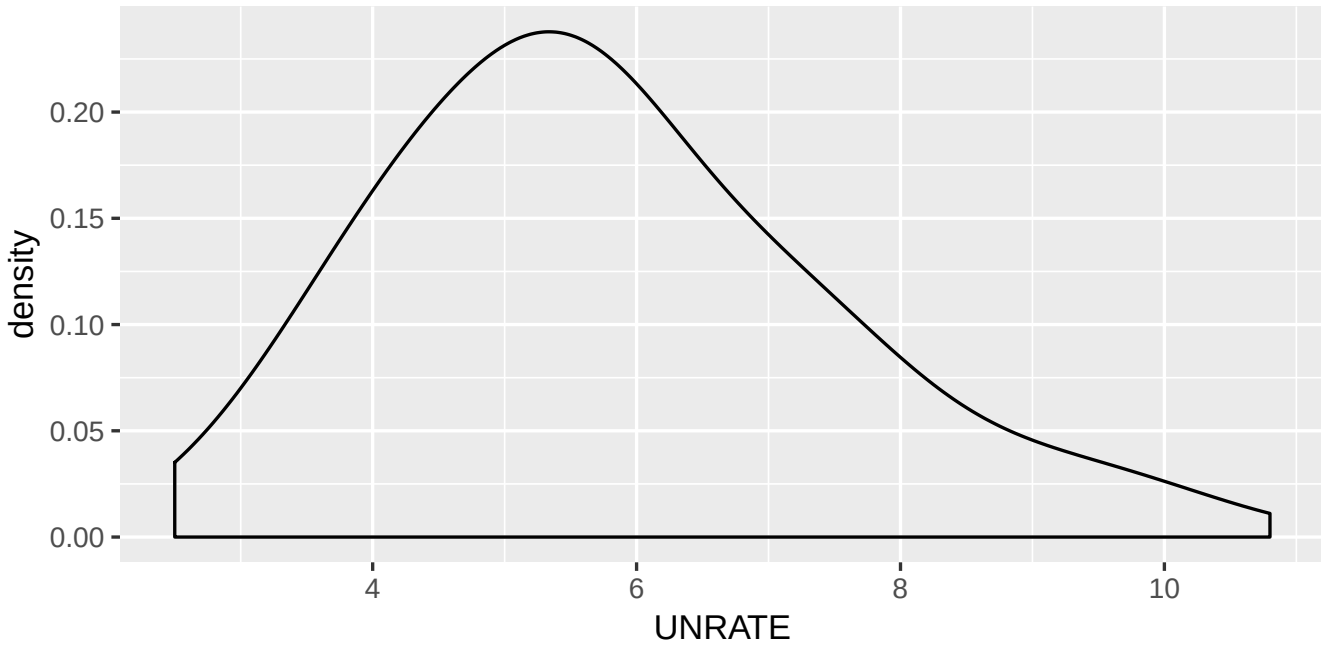
Two Properties of a Probability Density Function

1. It is non-negative; $f(x) \geq 0$. It can be greater than 1.
2. It integrates to one.

Lets visualize the relationship between CDF and PDF. on See Theory website at Brown University.

Chapter 3: Probability Distributions (Discrete and Continuous)

```
  ggplot(data = UNRATE, aes(x=UNRATE)) +
    geom_density(adjust = 2)
```



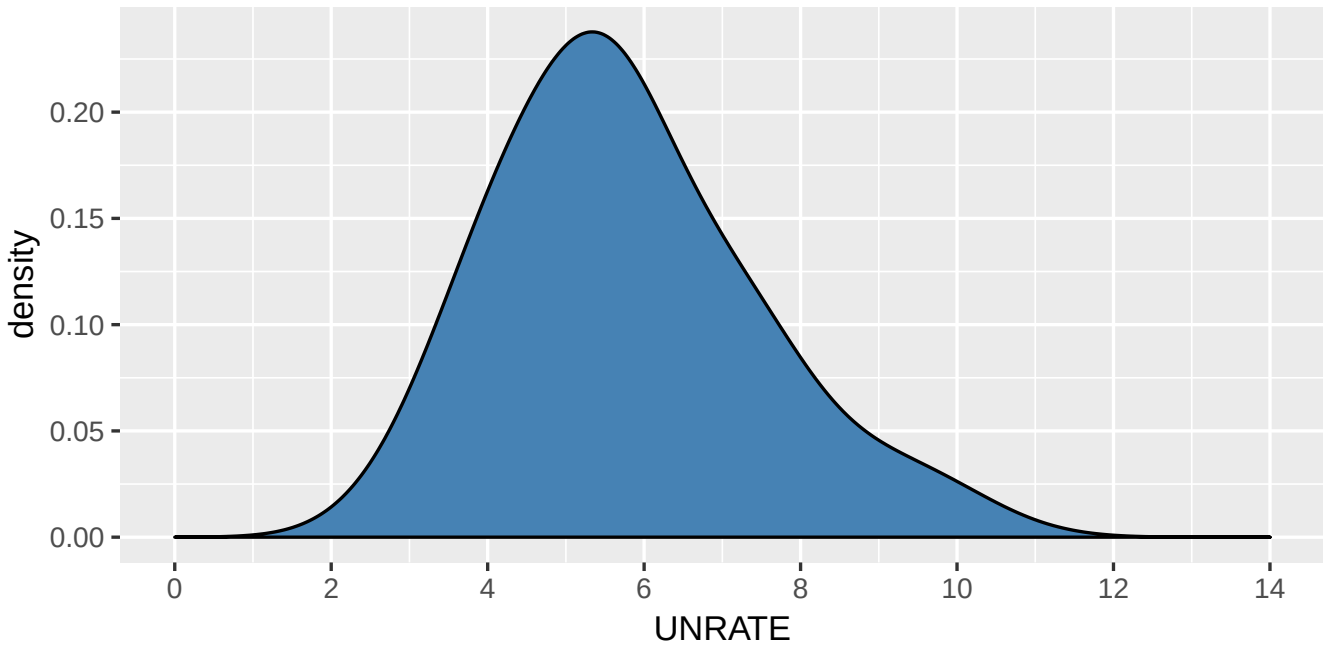

```
    expand_limits(x=c(0,14))


## mapping: x = ~x
## geom_blank: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

```
ggplot(data = UNRATE, aes(x=UNRATE)) +
  geom_density(adjust = 2)
  expand_limits(x=c(0,14))
```

1. bw: The smoothing bandwidth to be used. If numeric, the standard deviation of the smoothing kernel.
2. adjust: A multiplicate bandwidth adjustment.

The actual bandwidth is the product of these two values.

```
ggplot(data = UNRATE, aes(x=UNRATE)) +
  geom_density(fill = 'steelblue', adjust = 2) +
  expand_limits(x=c(0,14)) +
  scale_x_continuous(breaks=seq(0, 14, 2))
```



In the case of continuous variables, it is often easier to work with the CDF, but some of the features will be more visiable when using the PDF after we introduce other concepts.

CDF provides an way to evaluate all possible values. For all the possible values, y, we can calculate the probability of all the values being less than or equal to y.

A Degression: Kernel Density Estimation

How do we **actually** calculate the probability density function?

Kernel Density Estimation

Basis for many things in the future in machine learning, for example, classification

Recall that the definition of the derivative of a function $g(x)$ is given by

$$\frac{d}{dx}g(x) = \lim_{h \to 0} \frac{g(x+h) - g(x)}{h}$$

or, equivalently,

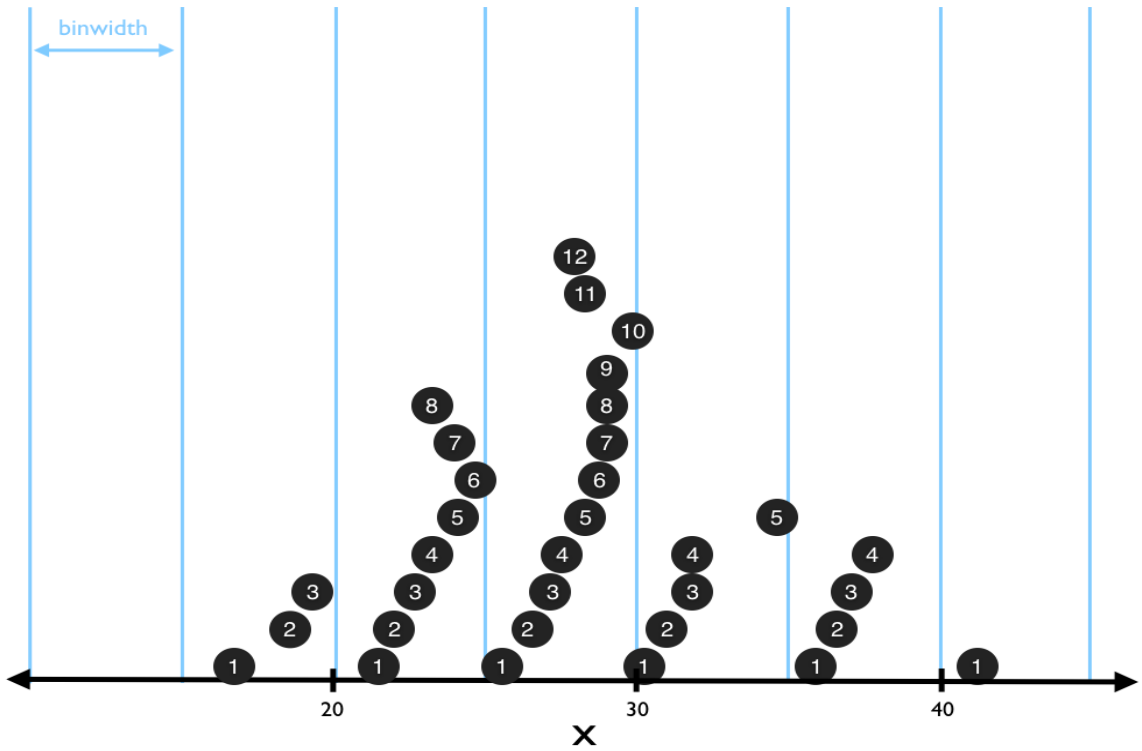
$$\frac{d}{dx}g(x) = \lim_{h \to 0} \frac{g(x+h) - g(x-h)}{2h}$$

$$\frac{d}{dx}g(x) = \lim_{h \to 0} \frac{g(x+h) - g(x-h)}{2h}$$

PDF is then given by

$$f(x) = \frac{d}{dx}F(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

$$f(x) = \frac{d}{dx}F(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$



$$\widehat{f}(x) = \frac{1}{2h} \cdot \frac{1}{N} \#\{X_1, X_2, \ldots, X_N \text{ falling in the interval } [x-h, x+h]\}$$

$$\widehat{f}(x) = \frac{1}{2h} \cdot \frac{1}{N} \sum \mathbb{I}\left[\{X_i \text{ falling in the interval } [x-h, x+h]\}\right]$$

Which X_i would fall in the interval $[x-h, x+h]$?

$$\widehat{f}(x) = \frac{1}{2h} \cdot \frac{1}{N} \#\{X_1, X_2, \ldots, X_N \text{ falling in the interval } [x-h, x+h]\}$$

$$\widehat{f}(x) = \frac{1}{2h} \cdot \frac{1}{N} \sum \mathbb{I}\left[\{X_i \text{ falling in the interval } [x-h, x+h]\}\right]$$

Which X_i would fall in the interval $[x-h, x+h]$?

$$\begin{aligned} x - h &\leq X_i \leq x + h \\ -h &\leq X_i - x \leq h \\ -1 &\qquad \frac{X_i - x}{h} \leq 1 \\ \left| \frac{X_i - x}{h} \right| &\leq 1 \end{aligned}$$

$$\widehat{f}(x) = \frac{1}{2h} \cdot \frac{1}{N} \sum \mathbb{I}\left[\{X_i \text{ falling in the interval } [x-h, x+h]\}\right]$$

is equivalent to

$$\widehat{f}(x) \quad = \quad \frac{1}{Nh} \sum \underbrace{\frac{1}{2} \cdot \mathbb{I}\left[\frac{X_i - x}{h} \mid \leq 1\right]}_{k(\frac{X_i - x}{h})}$$

$$\widehat{f}(x) = \frac{1}{2h} \cdot \frac{1}{N} \sum \mathbb{I}\left[\{X_i \text{ falling in the interval } [x-h, x+h]\}\right]$$

is equivalent to

$$\begin{aligned}
\widehat{f}(x) &= \frac{1}{Nh} \sum \underbrace{\frac{1}{2} \cdot \mathbb{I}\left[\frac{X_i - x}{h} \leq 1\right]}_{k(\frac{X_i - x}{h})} \\
\widehat{f}(x) &= \frac{1}{Nh} \sum k(\frac{X_i - x}{h})
\end{aligned}$$

where

$$k(\frac{X_i - x}{h}) = \begin{cases} \frac{1}{2} & \text{if } \frac{X_i - x}{h} \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

In practice, $k(v)$ can take many different parametric forms, e.g., standard normal kernel

$$k(v) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}v^2\}$$

Let's look at how to program this in R.

1. CDF $\rightarrow$ PDF (its derivative)
2. CDF $\rightarrow$ Quantile (its inverse function)

Quantile or Percentile

Quantile: τ-th Percentile (or Quantile)

$Q_\tau = \inf\{x : F(x) \geq \frac{\tau}{100}\}$

Quantile or Percentile

Quantile: τ-th Percentile (or Quantile)

$Q_\tau = \inf\{x : F(x) \geq \frac{\tau}{100}\}$

Loosely speaking, the τ-th percentile is rank.

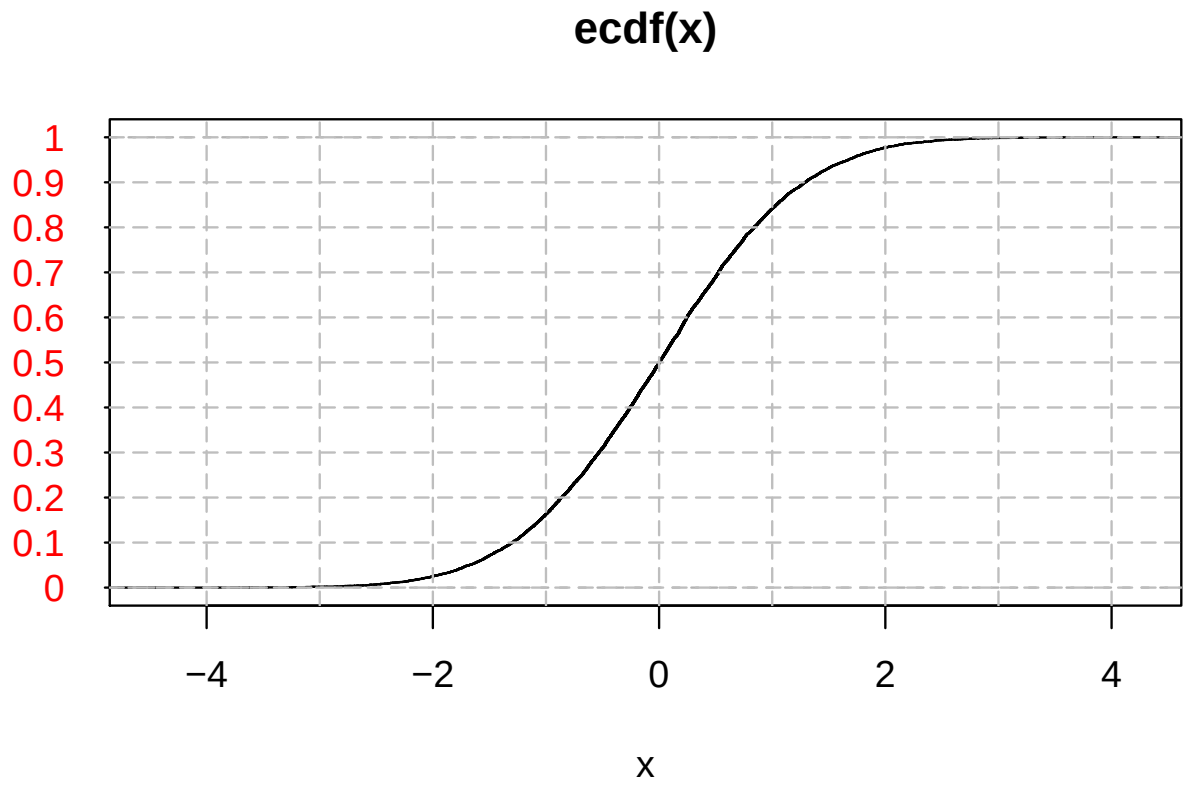
Quantile or Percentile

Quantile: τ-th Percentile (or Quantile)

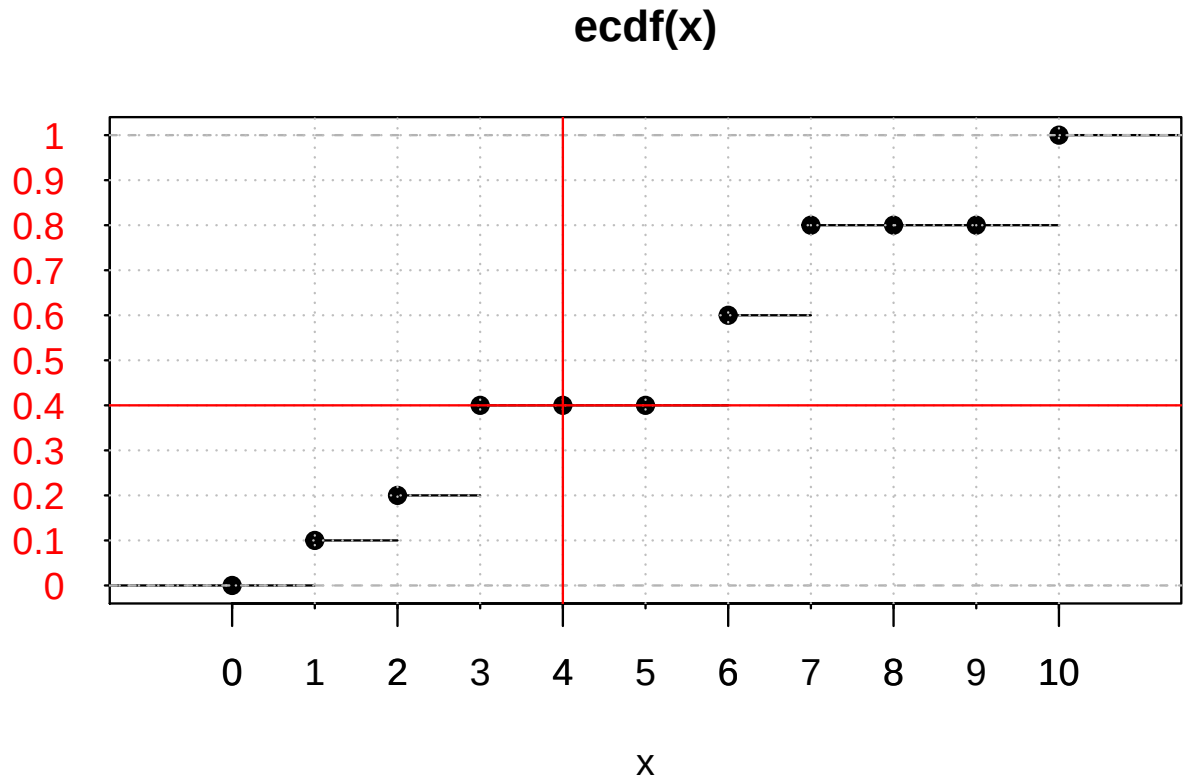
$Q_\tau = \inf\{x : F(x) \geq \frac{\tau}{100}\}$

Loosely speaking, the τ-th percentile is rank.

For example, 50^{th} percentile (median) indicates that fifty percent of the population is smaller than this value.


ecdf(x)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−4
−2
0
2
4
x

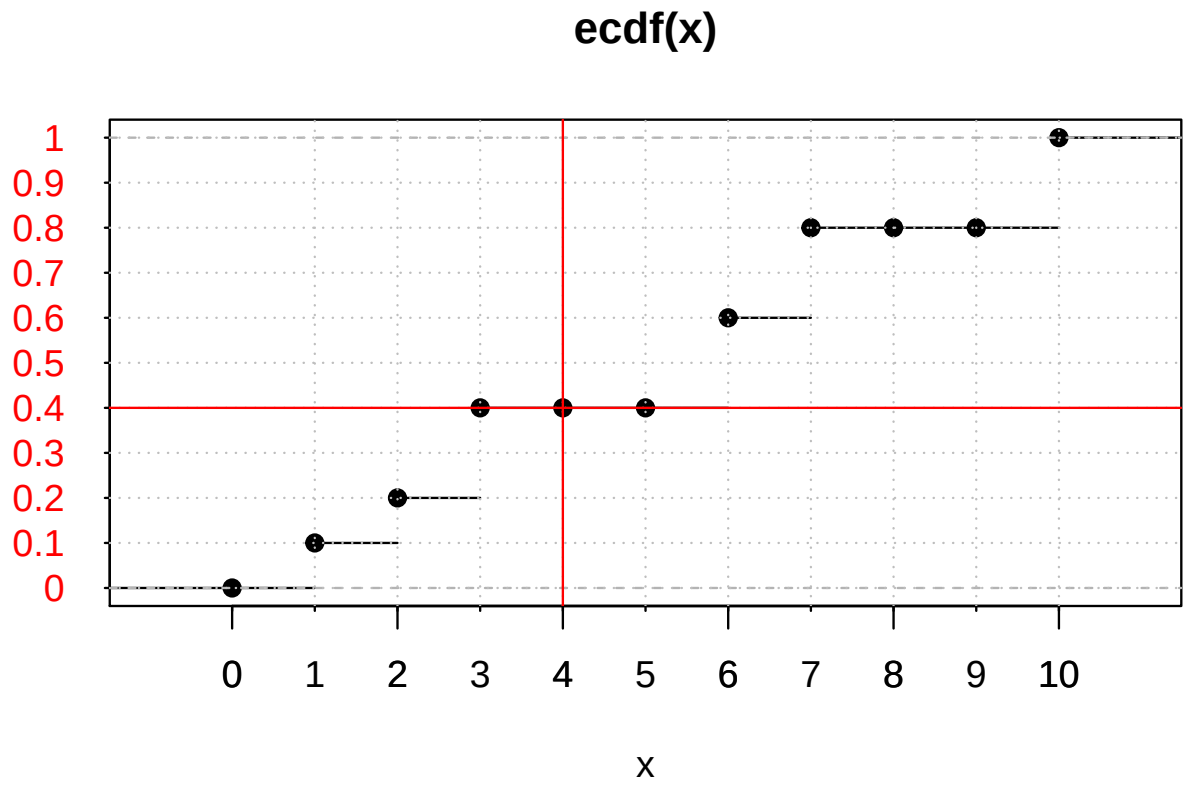
In Practice: What is your 40^{th} percentile in this dataset?



Quantile or Percentile

Quantile: τ-th Percentile (or Quantile)

$Q_\tau = \inf\{x : F(x) \geq \frac{\tau}{100}\}$

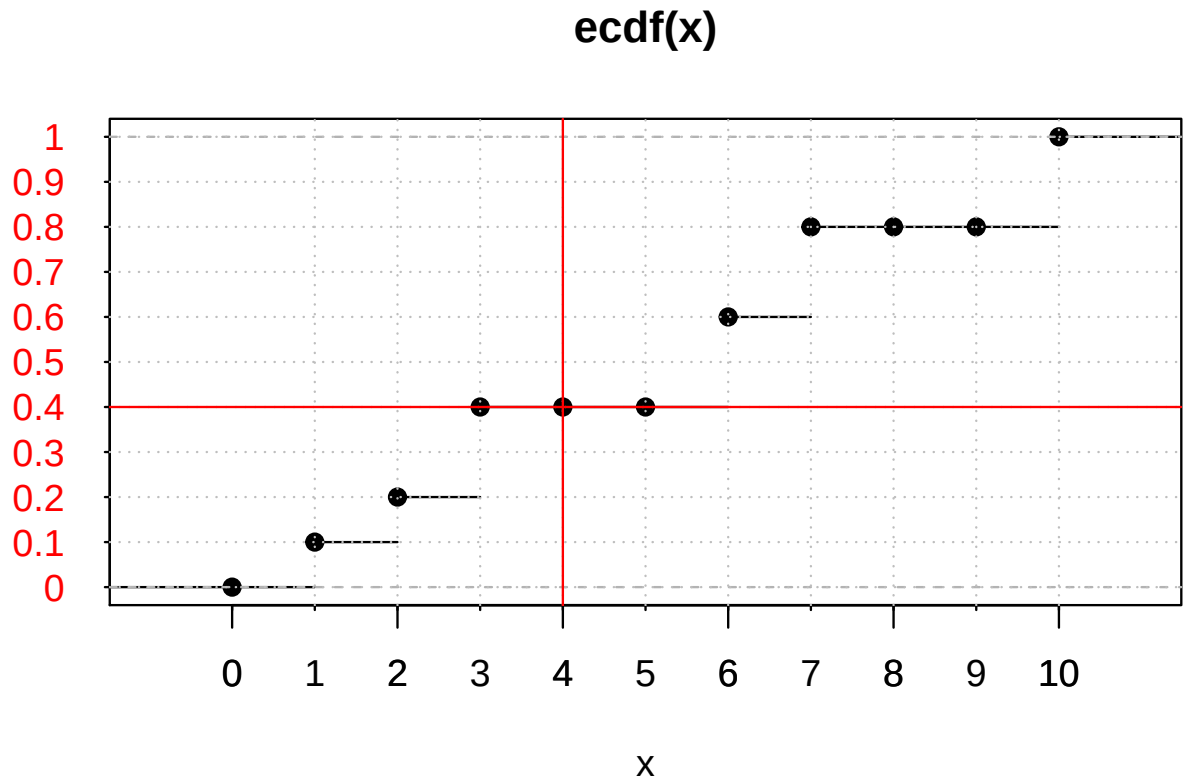


What are the xs with $F(x) \geq .4$? $(3, 4, 5, 6, 7, 8, 9, 10)$

In R, we can use the following command

```
x<-c(6,7,6,3,10,3,2,1,10,7)
median(x)

## [1] 6

quantile(x,.4)

## 40%
## 4.8
```

People have different definitions of quantiles when it is not unique!




```
quantile(x,.8, type = 1)

## 80%
##   7
```

This method is **nonparametric**, because we do not assume anything about the underlying distribution.

Application in Risk Management: **Value at Risk (VaR)** for measuring Downside Risk

Motivation:

1. Stock prices soar and plummet,
2. Interest rates escalate and drop,
3. Exchange rats rise and fall
4. etc..

Questions:

1. How low the asset or portfolio value could become?
2. How much loss could be made, tomorrow, next week, next month, next year?

I'd like to emphasize that we do not consider how to set up a porfolio, but consider an asset or when you already choose a portfolio.

In the context of constructing a portfolio, we are less concerned with individual risk and more concerned with the way the assets in the portfolio move together.

"Itineary"

1. Defintion of *VaR*
2. Worked Example for VaR for daily returns
3. VaR for alternative time horizons and rules to scale daily VaR
4. Alternative definition of returns
5. Cumulative risk profile
6. Expected shortfall measures
7. Risk comparisons, Interquantile Range, Boxplot

Value at Risk (VaR)

A statistical measure, against which the chance of a **worse** scenario happening is small (5 or 1 percent).

In its most general form, the Value at Risk measures the potential loss in value of a risky asset or portfolio over **a defined period** for **a given confidence interval**.

Warning:

Possible loss in value from "normal market risk", as opposed to all risk!

Example: VaR on an asset is $ 17.8 million at a one-week, 95% confidence level.

Example: VaR on an asset is $ 17.8 million at a one-week, 95% confidence level.

Translation:

Only a 5% chance that the value of the asset will drop more than $ 17.8 million over any given week.

VaR = Amount of Investment $\times$ VaR for Returns over a certain period

1. Amount of Investment
2. Confidence Interval (Percentage of Rare Events)
3. Returns and its distribution over a certain period

A Worked Example

VaR = Amount of Investment $\times$ VaR for Returns
over a certain period

1. Amount of Investment: $ 1000
2. Confidence Interval (Percentage of Rare Events): 5 percent
3. Returns and its distribution over a certain period

$$\text{VaR} = 1000 \times \text{VaR for Returns } \underline{\text{over a certain period}}$$

```r
# Set Parameters
# an investment of value V
V = 1000
alpha<-0.05
```

```r
# Load the package first before we can use those commands
library("quantmod")

# Download Walmart stock data from Yahoo
getSymbols('WMT')
```

```
## [1] "WMT"
```

```r
# Take a look at the data (first six rows/dates)
head(WMT)
```

```
##            WMT.Open WMT.High WMT.Low WMT.Close WMT.Volume WMT.Adjusted
## 2007-01-03    47.09    48.30   47.06     47.55   35687300     35.92604
## 2007-01-04    47.80    47.99   47.32     47.78   17073000     36.09982
## 2007-01-05    47.50    47.80   47.15     47.39   13556900     35.80516
## 2007-01-08    46.91    47.31   46.90     47.00   16396400     35.51049
## 2007-01-09    47.00    47.67   47.00     47.39   14643200     35.80516
## 2007-01-10    47.05    47.62   46.51     47.28   13348100     35.72205
```

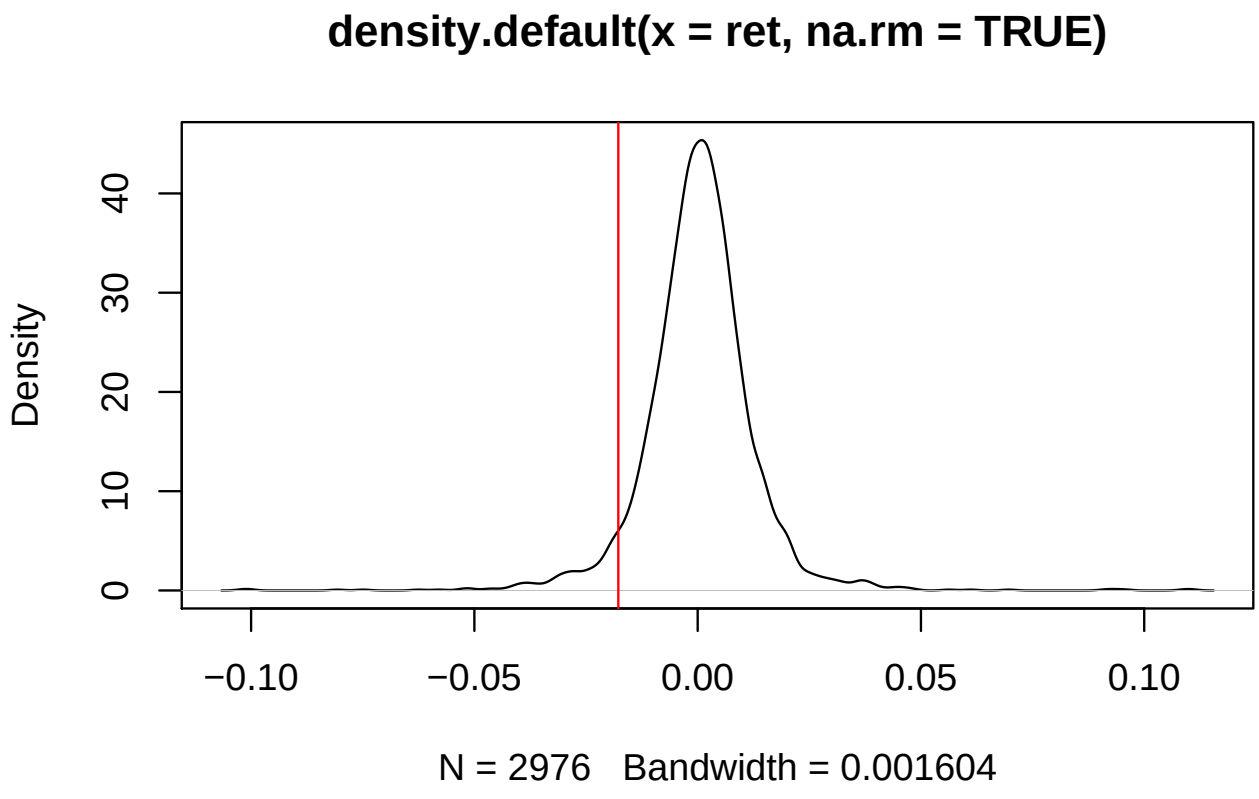
```r
# Calculate Daily Return
# return = log(S[i]/S[i-1])

#ret <- diff(log(WMT£WMT.Adjusted))
ret <- diff(WMT$WMT.Adjusted)/lag(WMT$WMT.Adjusted)
head(ret)
```

```
##            WMT.Adjusted
## 2007-01-03           NA
## 2007-01-04  0.004837020
## 2007-01-05 -0.008162423
## 2007-01-08 -0.008229736
## 2007-01-09  0.008298027
## 2007-01-10 -0.002321202
```

```r
ret2 <- dailyReturn(WMT$WMT.Adjusted)
head(ret2)
```

```
##            daily.returns
## 2007-01-03   0.000000000
## 2007-01-04   0.004837020
## 2007-01-05  -0.008162423
## 2007-01-08  -0.008229736
## 2007-01-09   0.008298027
## 2007-01-10  -0.002321202
```

Let's look at the VaR for daily returns



```r
quantile(ret,0.05, type = 1,na.rm=TRUE)
```

```
##           5%
## -0.01776873
```

```r
quantile(ret,alpha, type = 1,na.rm=TRUE)
```

```
##           5%
## -0.01776873
```

$$
\begin{aligned}
VaR &= \text{Amount of Investment} \times \text{ VaR for Daily Returns} \\
&= 1000 \times -0.0177687 \\
&= -17.7687333
\end{aligned}
$$

$$
\begin{aligned}
VaR &= \text{Amount of Investment} \times \text{ VaR for Daily Returns} \\
&= 1000 \times -0.0177687 \\
&= -17.7687333
\end{aligned}
$$

Interpretation: there is a 5% chance that we will lose more than $ -17.7687333 on our $1000 investment in the next day.

Our approach is nonparametric in that we do not impose any distributional assumptions in estimating the VaR for returns (**Historical VaR**).

This is different from traditional VaR analysis (JP Morgan/Reuters Metrics) which relies on the assumption of normal distribution (perhaps more precisely estimated, **Gaussian VaR**).

You can also choose a subset of the data (for example, more recent ones)

```
WMT.2011<-subset(ret2,index(ret2) >="2011-01-01" & index(ret2) <="2011-12-31")
head(WMT.2011)
```

```
##            daily.returns
## 2011-01-03   0.011681960
## 2011-01-04   0.003848821
## 2011-01-05  -0.006572918
## 2011-01-06  -0.008270498
## 2011-01-07   0.002223665
## 2011-01-10  -0.006471820
```

```
tail(WMT.2011)
```

```
##            daily.returns
## 2011-12-22  -0.003367581
## 2011-12-23   0.013515822
## 2011-12-27  -0.002667080
## 2011-12-28  -0.001671527
## 2011-12-29   0.004353016
## 2011-12-30  -0.003834163
```

You can also calculate VaR for different time horizone, for example, for weekly returns. The only thing you need to do is to use appropriate commands to calculate the corresponding returns. For example,

```
WMT.week<-weeklyReturn(WMT$WMT.Adjusted)
WMT.month<-monthlyReturn(WMT$WMT.Adjusted)

head(WMT.week)
```

```
##            weekly.returns
## 2007-01-05   -0.003364885
## 2007-01-12    0.012449883
## 2007-01-19    0.006878169
## 2007-01-26   -0.013247985
## 2007-02-02    0.008600591
## 2007-02-09   -0.002287450
```

In industry, people also use the Square Root of T rule to scale 1-day VaR to a $T - day$ VaR

Assuming that the returns are independent and ientically distributed, we can calculate a T-day VaR by multiplying the daily VaR by $\sqrt{T}$. For example,

our VaR for daily returns is -17.5321817, and then 7day VaR should be $-17.5321817 \times \sqrt{7} = -46.38579$

You can also calculate VaR for returns using log formula.

$$\log(\text{Price}_t) - \log(\text{Price}_{t-1})$$

```
WMT.log<-dailyReturn(WMT$WMT.Adjusted, type='log')
head(WMT.log)
```

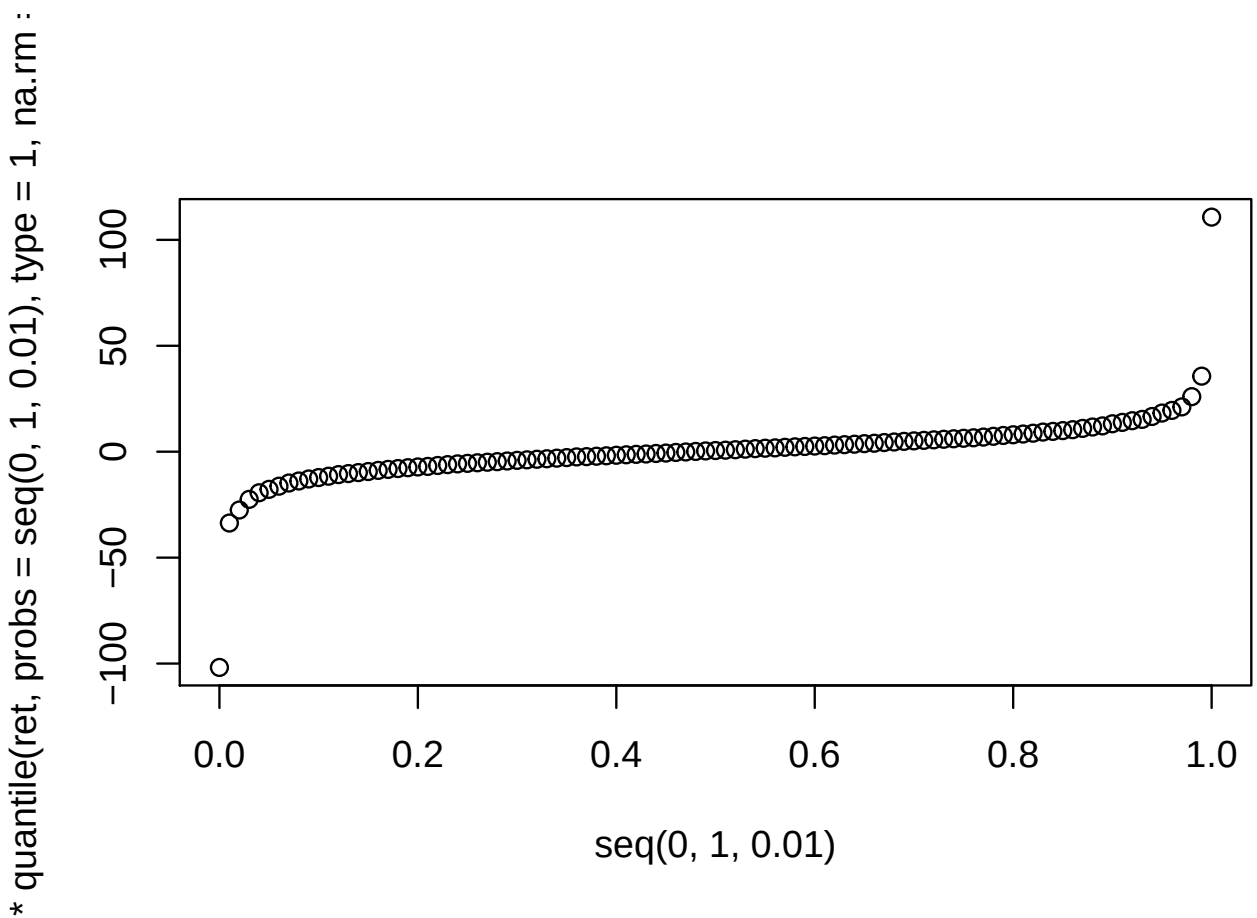
```
##            daily.returns
## 2007-01-03   0.000000000
## 2007-01-04   0.004825359
## 2007-01-05  -0.008195918
## 2007-01-08  -0.008263788
## 2007-01-09   0.008263788
## 2007-01-10  -0.002323900
```

```
head(ret2)
```

```
##            daily.returns
## 2007-01-03   0.000000000
## 2007-01-04   0.004837020
## 2007-01-05  -0.008162423
## 2007-01-08  -0.008229736
## 2007-01-09   0.008298027
## 2007-01-10  -0.002321202
```

Instead of calculate VaR for a particular level, we can also present the entire *cumulative risk profile*

```
plot(seq(0,1,0.01),1000*quantile(ret,probs = seq(0,1,0.01), type=1,na.rm=TRUE))
```



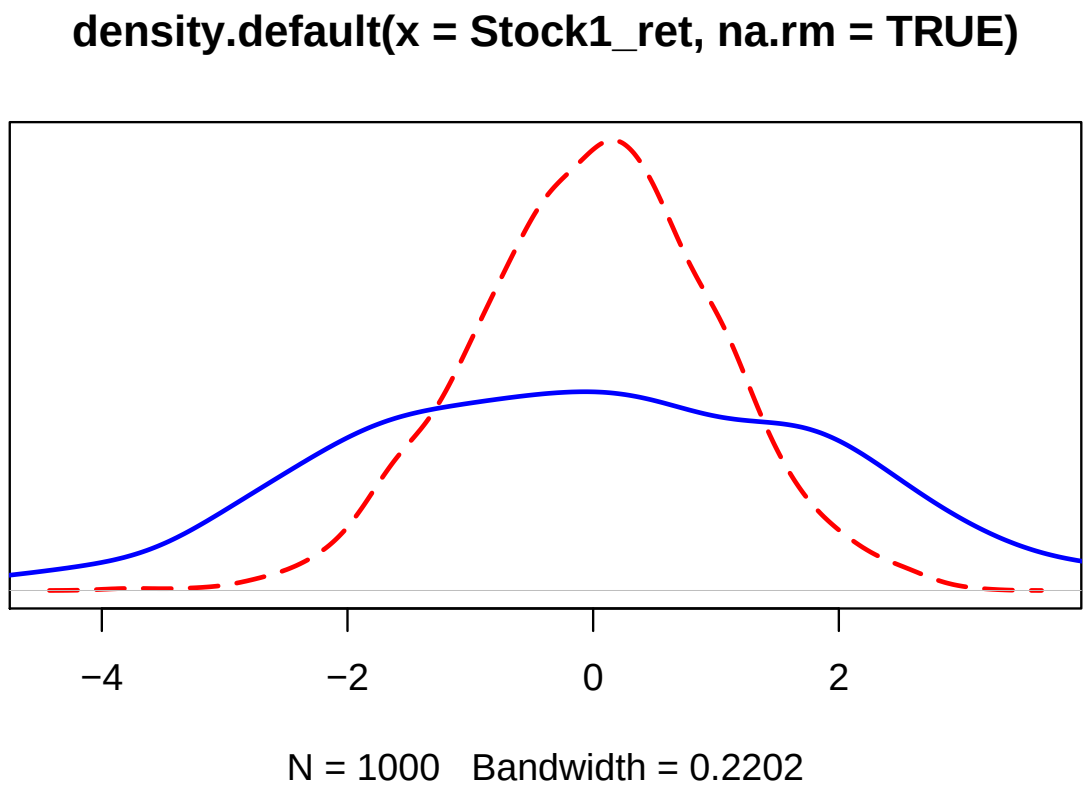
One issue with VaR is that it does not capture the sape of the tail of the distribution. That is VaR does not allow us to answer the question, "if losses exceed VaR, how bad should we expect that loss to be?"

Another commonly used measure is

Expected Shortfall: $\mathbb{E}[return|return \leq Q_\tau]$

where Q_τ is τ^{th} percentile.

Question: Which of the following stocks will have a larger VaR?



```r
quantile(Stock1_ret,0.05, type = 1)
```

```
##         5%
## -1.628805
```

```r
quantile(Stock2_ret,0.05, type = 1)
```

```
##         5%
## -3.136011
```

$$
\begin{aligned}
VaR &= \text{Amount of Investment} \times \text{ VaR for Daily Returns} \\
&= 1000 \times -1.6288051 \\
&= -1628.8051446
\end{aligned}
$$

$$
\begin{aligned}
VaR &= \text{Amount of Investment} \times \text{ VaR for Daily Returns} \\
&= 1000 \times -3.1360113 \\
&= -3136.0112578
\end{aligned}
$$

The previous example demonstrates one rule:

The more volatile (spread out) the rate of return, the larger is the VaR (the potential loss on your investment)!

Based on these percentiles, we can also define a measure of spread, called **Interquantile Range (IQR)**

$$75^{th} \text{ percentile - } 25^{th} \text{ percentile}$$

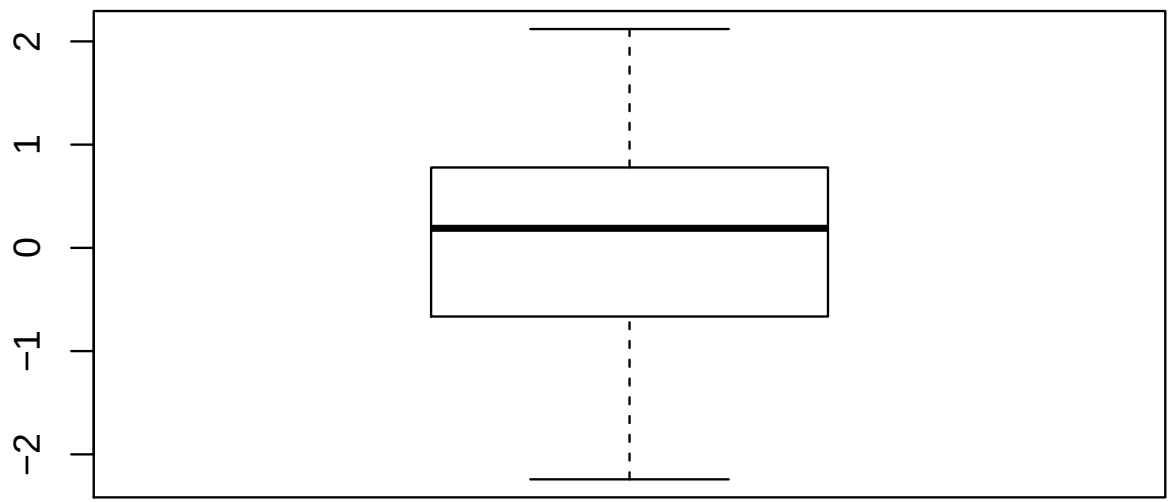
Another technique: **Box Plot**

```
##   [1]  0.79254435  1.79759848  1.38536369 -0.57259890 -1.09379987
##   [6]  0.22340317  1.56730123  2.11951719  0.05451596 -1.07671391
##  [11]  0.71704492  0.97085334  0.70225681  0.44420861  1.90614515
##  [16] -0.99192140  0.56162400 -0.76596153 -0.84160884  0.21774059
##  [21] -2.06077769 -0.32947375  1.83824488  1.58000658  0.56481325
##  [26] -0.22220182  0.02481441 -0.85487530  0.41383057  1.37976239
##  [31] -1.51515971  0.50674990 -1.20894738 -1.69894846 -1.75601452
##  [36]  0.54014995  0.61089722  0.23907275 -0.98405887  1.05962723
##  [41]  0.20435251  0.48693831  1.16288478 -0.99539551  1.61250261
##  [46]  0.76445547 -0.48752825  2.08002293  0.86762120 -1.04240113
##  [51] -0.39177042  0.99283739 -1.16612532  1.29317359  2.06502739
##  [56] -0.09205054 -0.55430761  0.17667481  1.64526054 -2.24213154
##  [61]  0.82194737 -0.16278638 -0.23591590 -0.58547299 -0.16278048
##  [66] -0.39282954 -0.12772479  0.91296668 -0.87075827 -0.83286930
##  [71] -0.93258122 -1.56267682  0.41724754 -0.45762519  0.26953909
##  [76] -0.36042887  0.39720737  1.08388208 -0.43528859 -0.17294339
##  [81]  1.18612936  1.24219280  0.74322767 -0.73462368  0.29641483
##  [86]  0.11322785 -0.59514616  0.08635186  0.28241537 -1.47421951
##  [91]  0.20210545  1.00008845  0.13306396 -1.19893323  0.30192690
##  [96]  0.74494305 -0.24501034  0.66312846 -1.05398803 -0.76848352
```

```
summary(box.data)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.2421 -0.6300  0.1894  0.1016  0.7715  2.1195
```

```
boxplot(box.data)
```

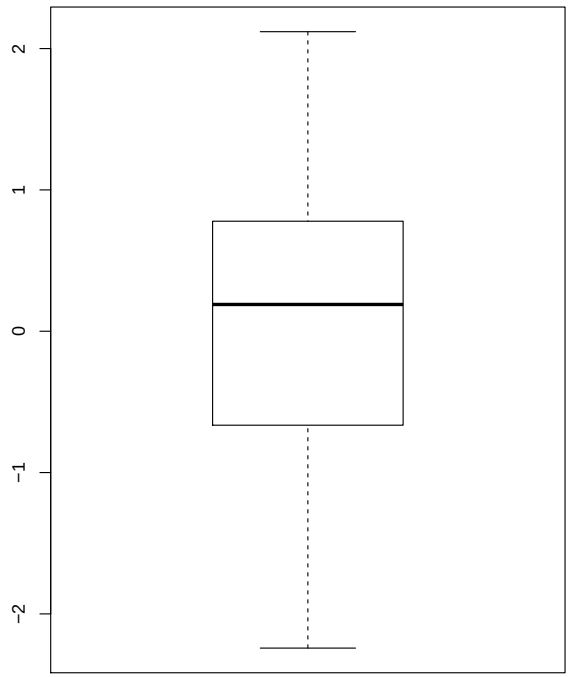


The boxplot compactly displays the distribution of a continuous variable. It visualises five summary statistics

1. the median
2. two hinges
3. two whiskers

Graph

```
boxplot(box.data)
```

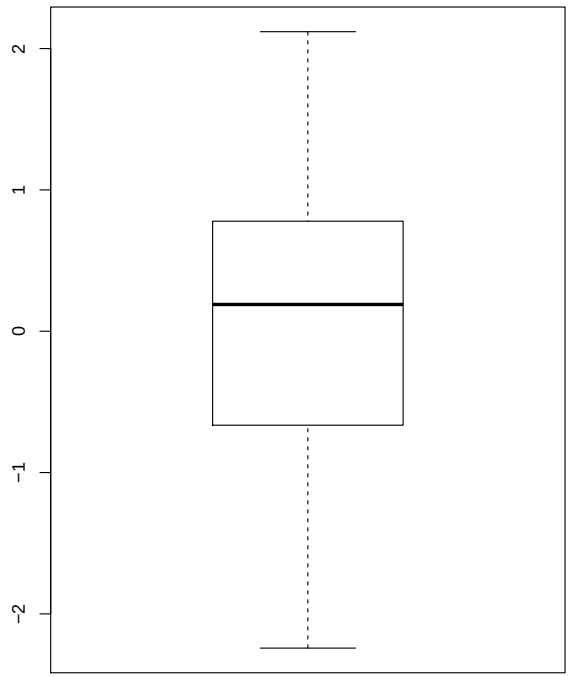


Interpretation

1. The top and bottom lines of the rectangle are the 3rd and 1st quartiles (Q3 and Q1), respectively. The length of the rectangle is the IQR.

Graph

```
boxplot(box.data)
```

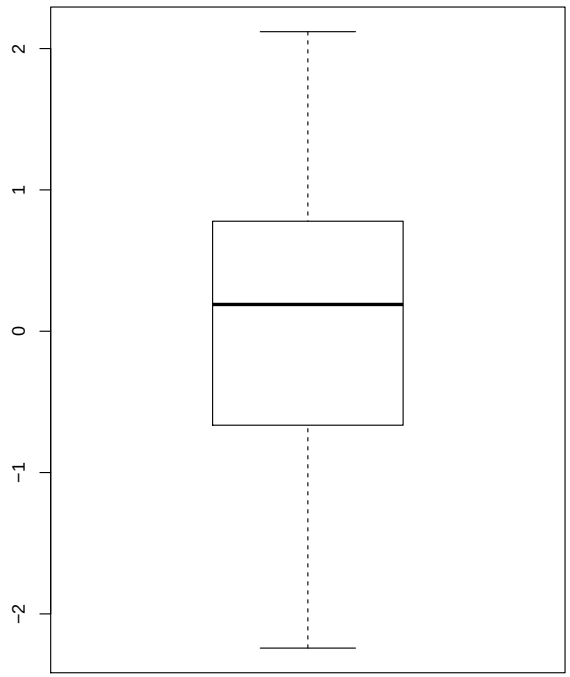


Interpretation

1. The top and bottom lines of the rectangle are the 3rd and 1st quartiles (Q3 and Q1), respectively. The length of the rectangle is the IQR.
2. The line in the middle of the rectangle is the median.

Graph

```
boxplot(box.data)
```

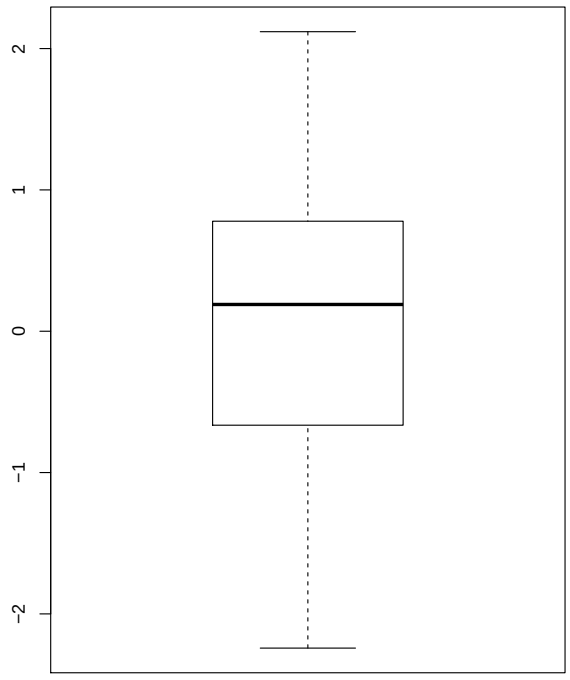


Interpretation

1. The top and bottom lines of the rectangle are the 3rd and 1st quartiles (Q3 and Q1), respectively. The length of the rectangle is the IQR.
2. The line in the middle of the rectangle is the median.
3. The top whisker denotes the maximum value or the 3rd quartile plus 1.5 times the interquartile range (Q3 + 1.5*IQR), whichever is smaller.

Graph

```
boxplot(box.data)
```



Interpretation

1. The top and bottom lines of the rectangle are the 3rd and 1st quartiles (Q3 and Q1), respectively. The length of the rectangle is the IQR.
2. The line in the middle of the rectangle is the median.
3. The top whisker denotes the maximum value or the 3rd quartile plus 1.5 times the interquartile range (Q3 + 1.5*IQR), whichever is smaller.
4. The bottom whisker denotes either the minimum value or the 1st quartile minus 1.5 times the interquartile range (Q1 - 1.5*IQR), whichever is