Empirical distribution

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom

Attaching package: 'gridExtra'


The following object is masked from 'package:dplyr':

    combine
```

Work with some penguin data

```
adelie <- penguins |>
    filter(`species` == "Adelie") |>
    drop_na()
```

Histogram and ECDF

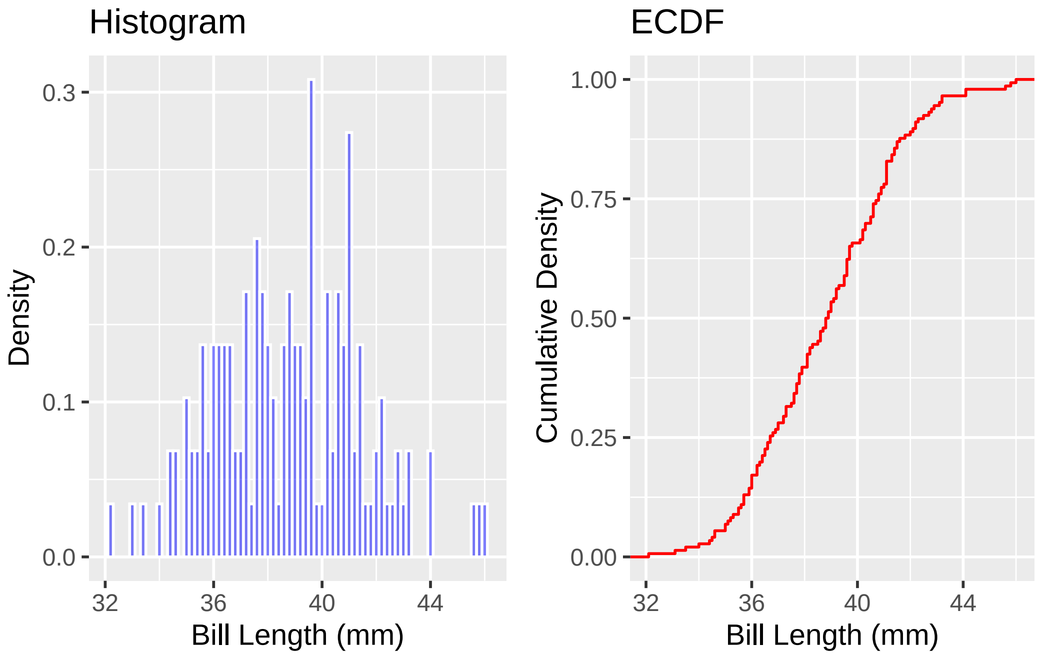
A *histogram* of a random variable gives the number of samples in ranges

$$a \leq x \leq a + u$$

Each range is sometimes called a bin. A histogram *approximates the probability density.*

The *empirical cumulative distribution* (ECDF) shows, for each value a of a random variable x, the fraction of sample points where $x \leq a$. The ECDF *approximates the cumulative distribution.*

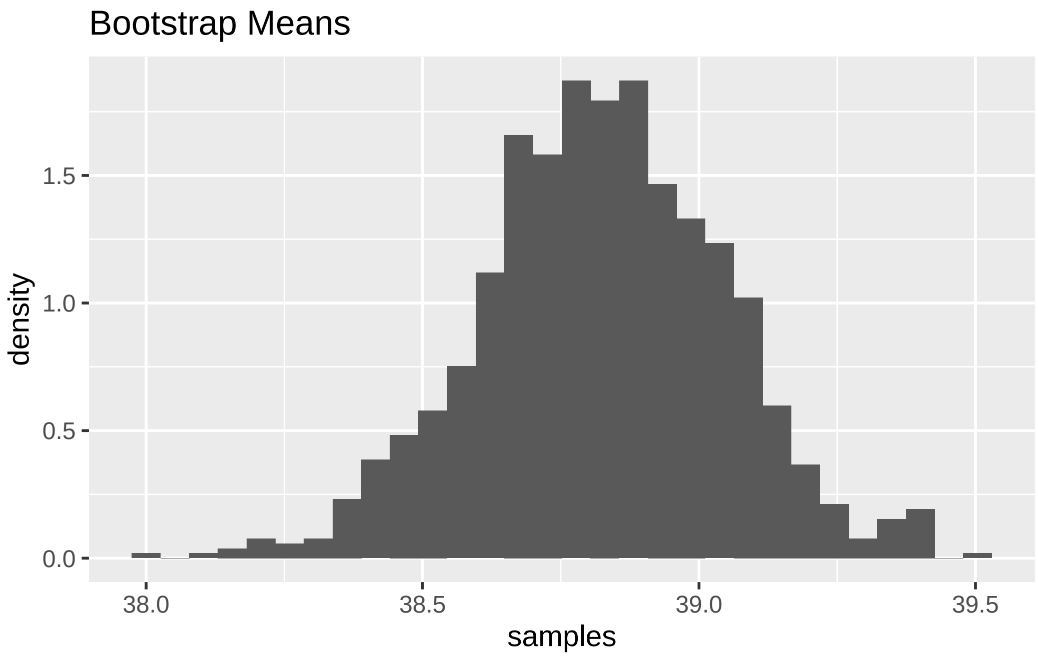
```
hist<-ggplot(data=adelie)+geom_histogram(aes(x=`bill_length_mm`,y=after_stat(density)),bin
labs(x="Bill Length (mm)",y="Density", title="Histogram")
ecdf <-ggplot(data=adelie)+stat_ecdf(aes(x=`bill_length_mm`),color='red')+labs(x="Bill Len
grid.arrange(hist, ecdf, ncol = 2)
```



Bootstrap means histogram

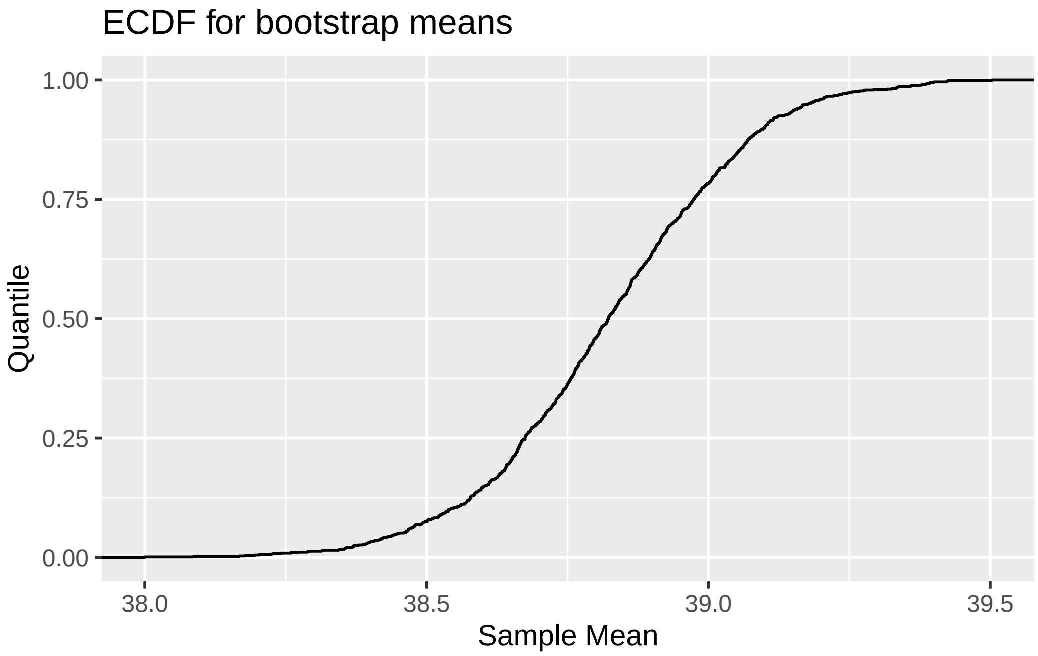
```
bootstrap_mean <- function(data) {
    sample <- data |> sample_frac(1,replace=TRUE)
    return(mean(pull(sample)))

}

samples<-replicate(1000,bootstrap_mean(adelie|>select("bill_length_mm")))
ggplot()+geom_histogram(aes(x=samples,y=after_stat(density)))+labs(title='Bootstrap Means'
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



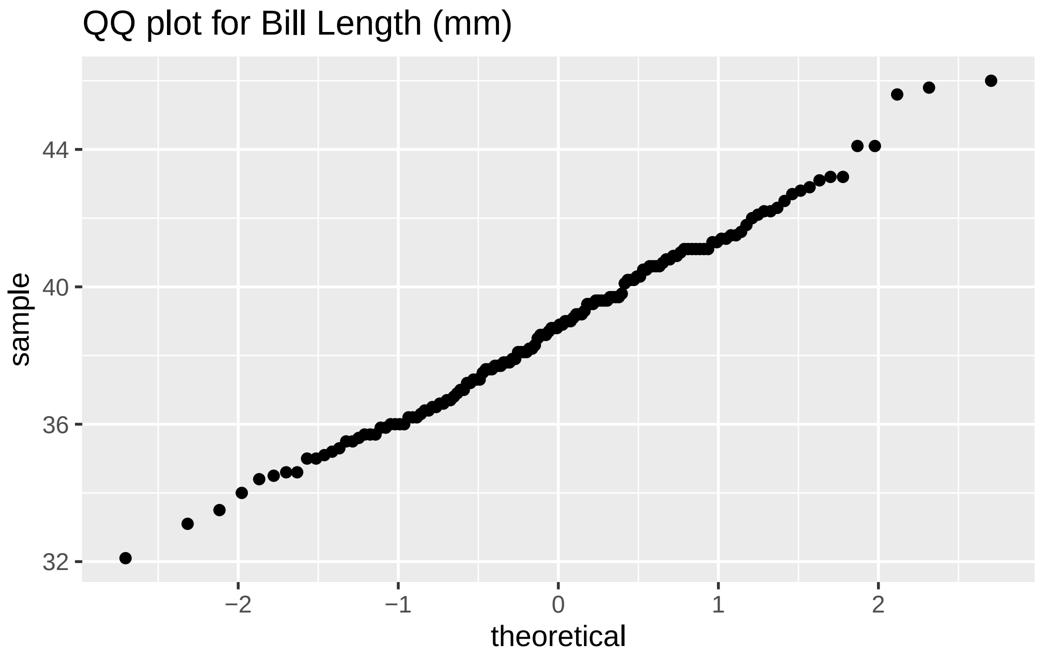
Bootstrap means ECDF

```
ggplot() + stat_ecdf(aes(x=samples))+labs(title="ECDF for bootstrap means",x="Sample Mean"
```

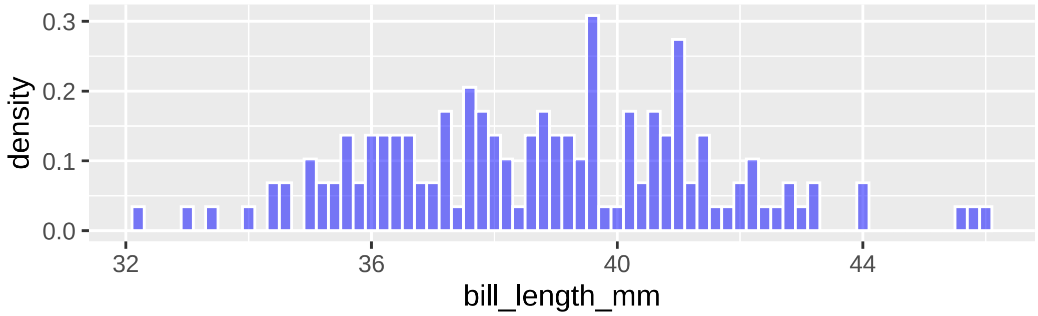


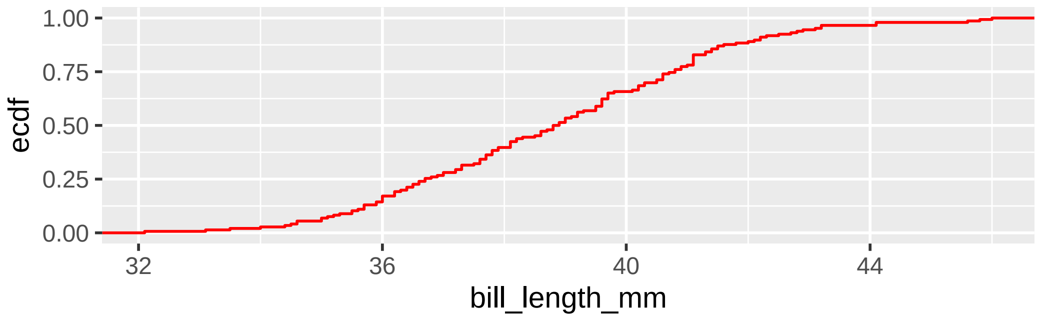
QQ plot

```r
ggplot(data=adelie)+stat_qq(aes(sample=`bill_length_mm`))+
    labs(title="QQ plot for Bill Length (mm)")
```



```r
show <- function(data, column) {
    hist<-ggplot(data=data)+geom_histogram(aes(x={{column}},y=after_stat(density)),binwidt
    ecdf <-ggplot(data=data)+stat_ecdf(aes(x={{column}}),color='red')
    return(list(hist,ecdf))
}
grid.arrange(grobs=show(adelie,bill_length_mm),ncol=1)
```


0.3
0.2
0.1
0.0
density
32
36
40
44
bill_length_mm


1.00
0.75
0.50
0.25
0.00
ecdf
32
36
40
44
bill_length_mm