ü6

Ben Gutzeit

2025-12-08

1.1 In your own words, describe what one row of the object bwt represents

- **low** = Low birth weight
- **age** = Mother's age in years
- **lwt** = Mother's weight at last menstrual period (in pounds)
- **race** = Mother's race (White, Black, or Other)
- **smoke** = Whether the mother smoked during pregnancy (Yes/No)
- **ptl** = Number of previous premature labors
- **ht** = History of hypertension (Yes/No)
- **ui** = Presence of uterine irritability (Yes/No)
- **ftv** = Number of physician visits during the first trimester
- **bwt** = Baby's birth weight in grams

1.2 Identify which variables are categorical and which are numeric.

```
df <- data.frame(
  categorical = c("low","race","smoke","ht","ui"),
  numeric = c("age","lwt","ptl","ftv","bwt")
)

df
```

```
##   categorical numeric
## 1         low     age
## 2        race     lwt
## 3       smoke     ptl
## 4          ht     ftv
## 5          ui     bwt
```

1.3 For each categorical variable, write down the meaning of its levels.

low

- Yes = low birth weight

- No = normal birth weight

race

- White = white skintone
- Black = black skintone
- Other = other skintone

smoke

- Yes = mother is smoker
- No = mother is no smoker

ht

- Yes = history of hypertension
- No = no history of hypertension

ui

- Yes = uterine irritability
- No = no uterine irritability

2.1 Create a bar chart for and low using ggplot2.

für was und low?

2.2 Briefly comment on the proportion of low birth weight (low = "Yes")

```
lyes <- sum(bwt$low == "Yes")
lno <- sum(bwt$low == "No")
lall <- sum(bwt$low != "")

lyes <- lyes / lall
lno <- lno / lall

ldf <- data.frame(
  yes = lyes,
  no = lno
  )

ldf
```

```
##         yes        no
## 1 0.3121693 0.6878307
```

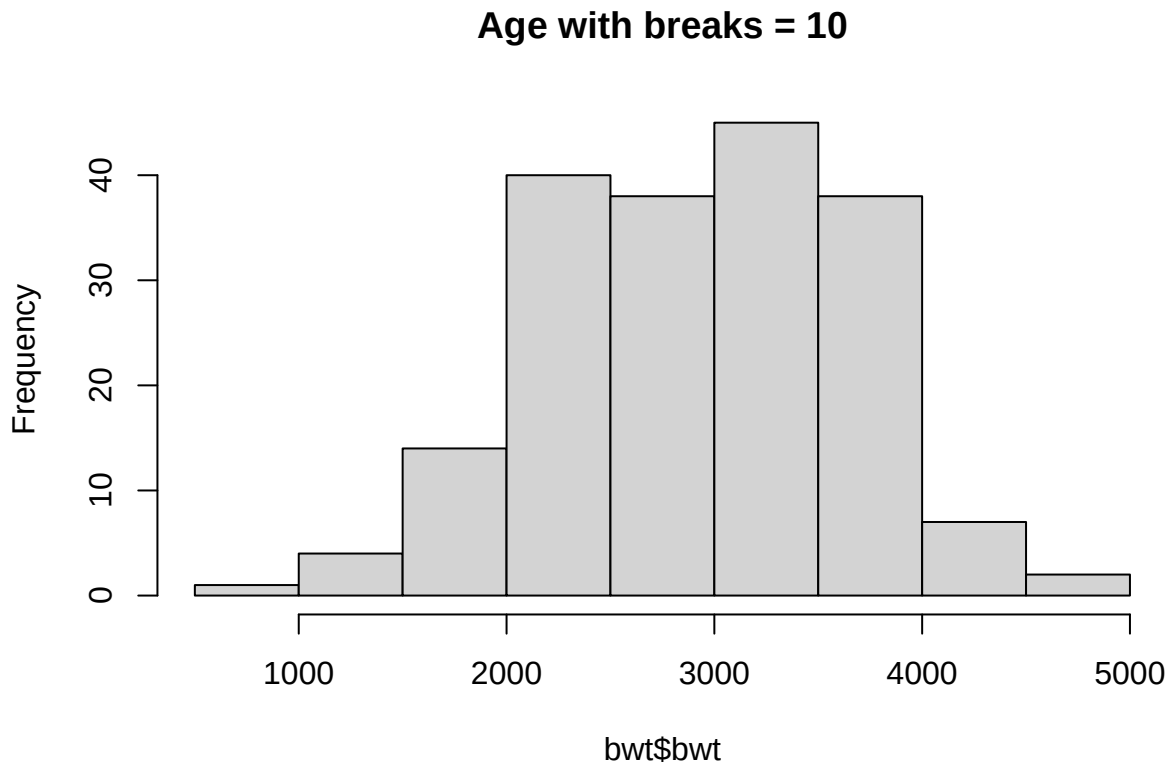
```r
#frequency table
table(bwt$low)
```

```
##
##  No Yes
## 130  59
```

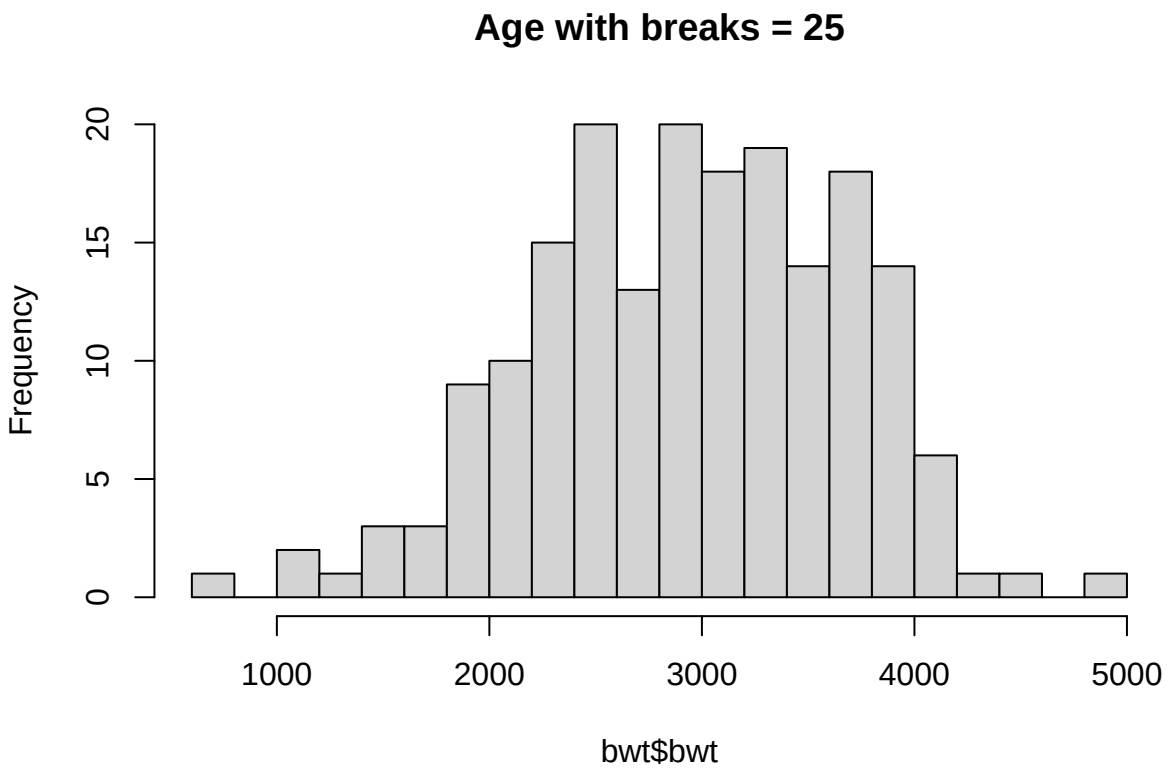
~ 31% of the babies are born with underweight

3.1 Produce histograms of bwt with three different bin widths, e.g. 100, 250, and 500.

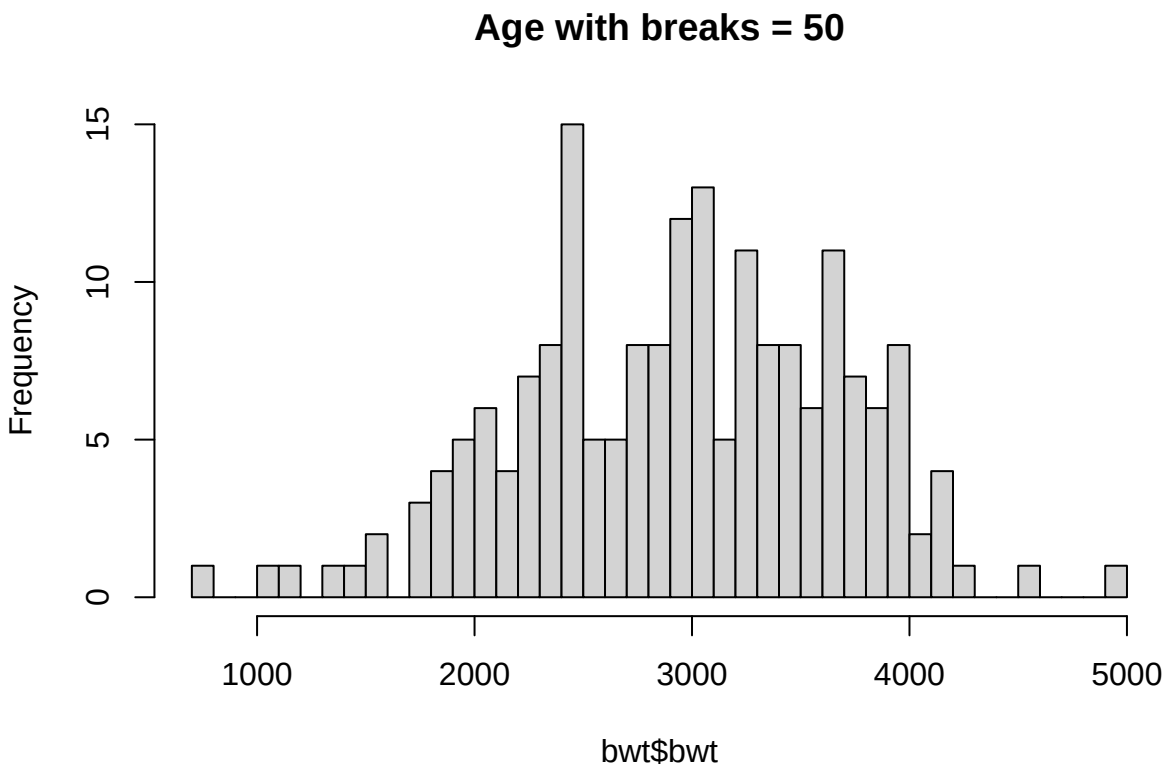
```r
hist(bwt$bwt, breaks = 10 , main="Age with breaks = 10")
```



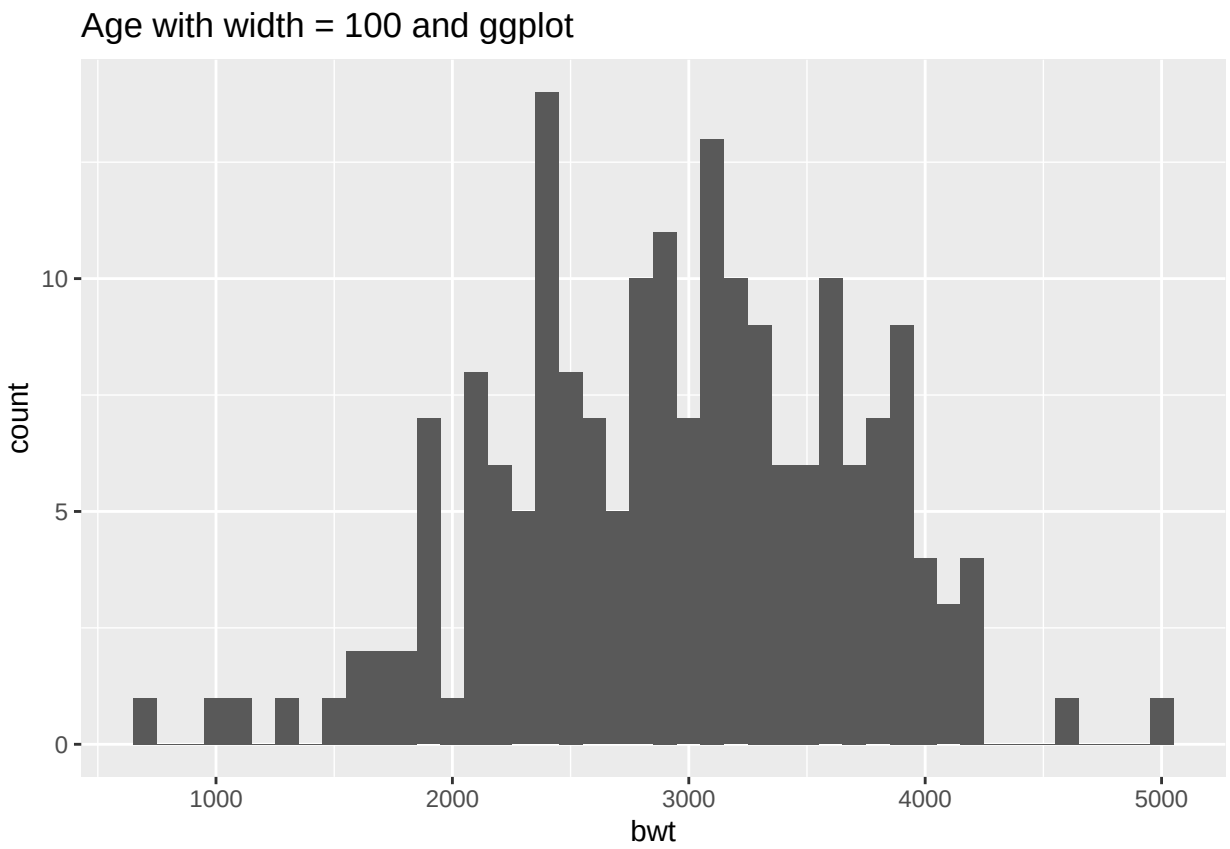

```r
hist(bwt$bwt, breaks = 25 , main="Age with breaks = 25")
```



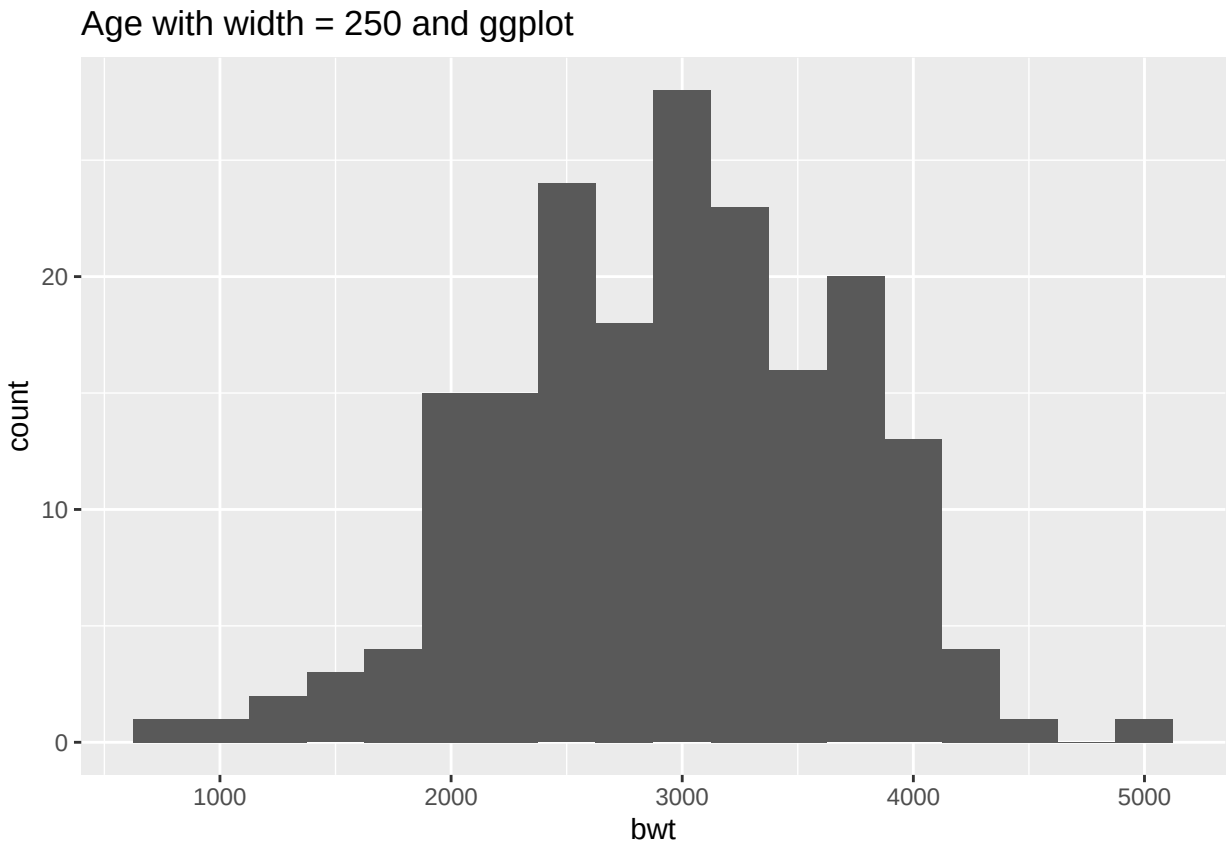

```r
hist(bwt$bwt, breaks = 50 , main="Age with breaks = 50")
```



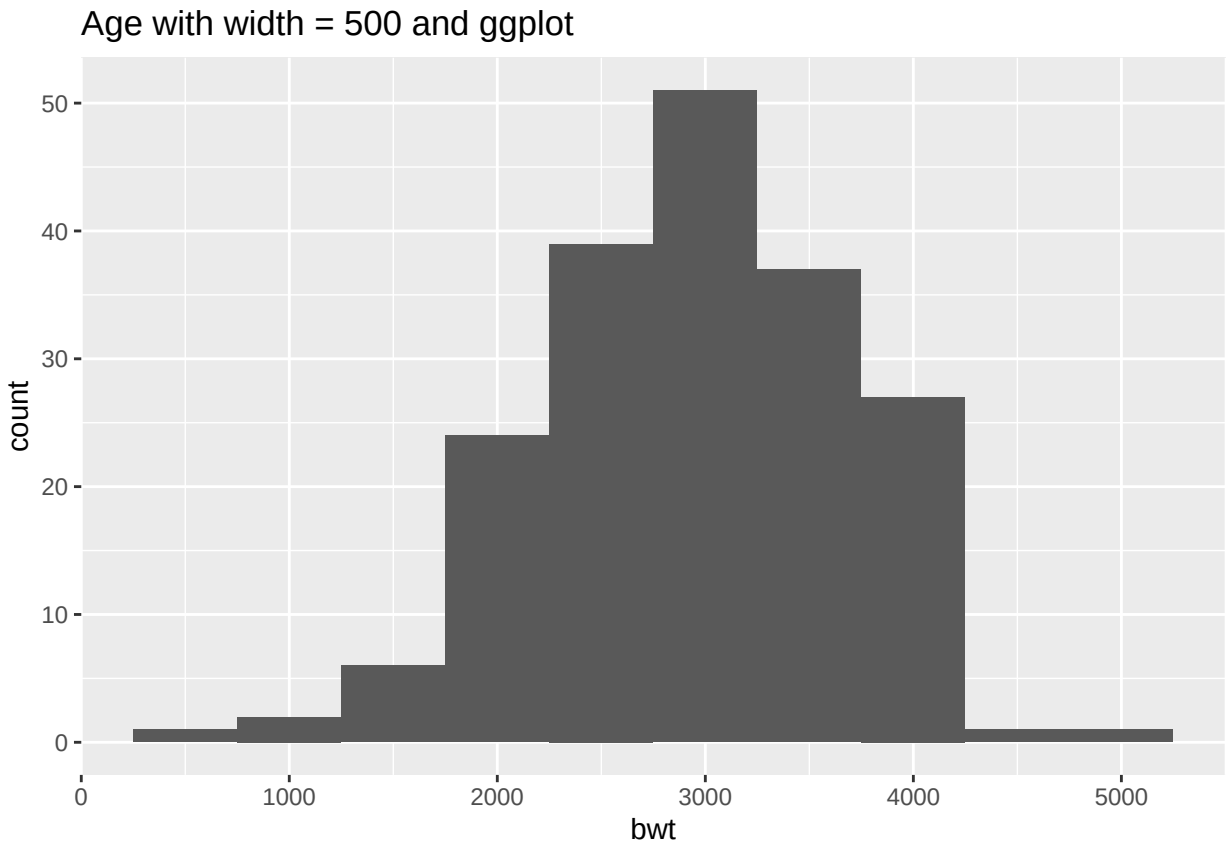

```
ggplot(bwt, aes(x = bwt)) +
  geom_histogram(binwidth = 100) +
  ggtitle("Age with width = 100 and ggplot")
```




```
ggplot(bwt, aes(x = bwt)) +
  geom_histogram(binwidth = 250)+
  ggtitle("Age with width = 250 and ggplot")
```




```
ggplot(bwt, aes(x = bwt)) +
  geom_histogram(binwidth = 500)+
  ggtitle("Age with width = 500 and ggplot")
```



3.2 Describe how the choice of bin width changes the appearance of the histogram.

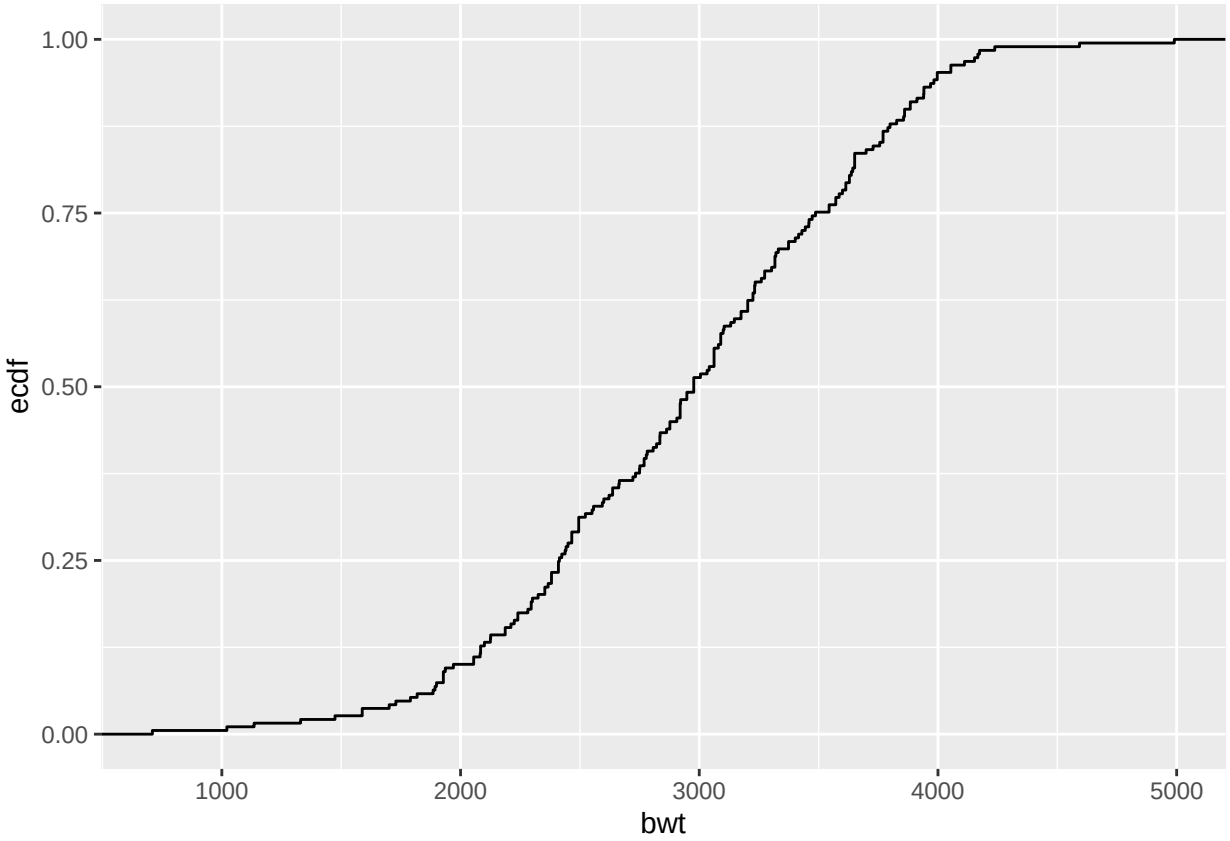
"breaks" changes the widths of the bars

3.3 Does the distribution of birth weight appear roughly symmetric, right-skewed, or leftskewed?

- roughly symmetric

4.1 Plot the empirical cumulative distribution function (ECDF) of bwt using stat_ecdf().

```
ggplot(bwt, aes(x = bwt)) + stat_ecdf()
```


1.00
0.75
0.50
0.25
0.00
ecdf
1000
2000
3000
4000
5000
bwt