

MATH 203: PRINCIPLES OF STATISTICS 1

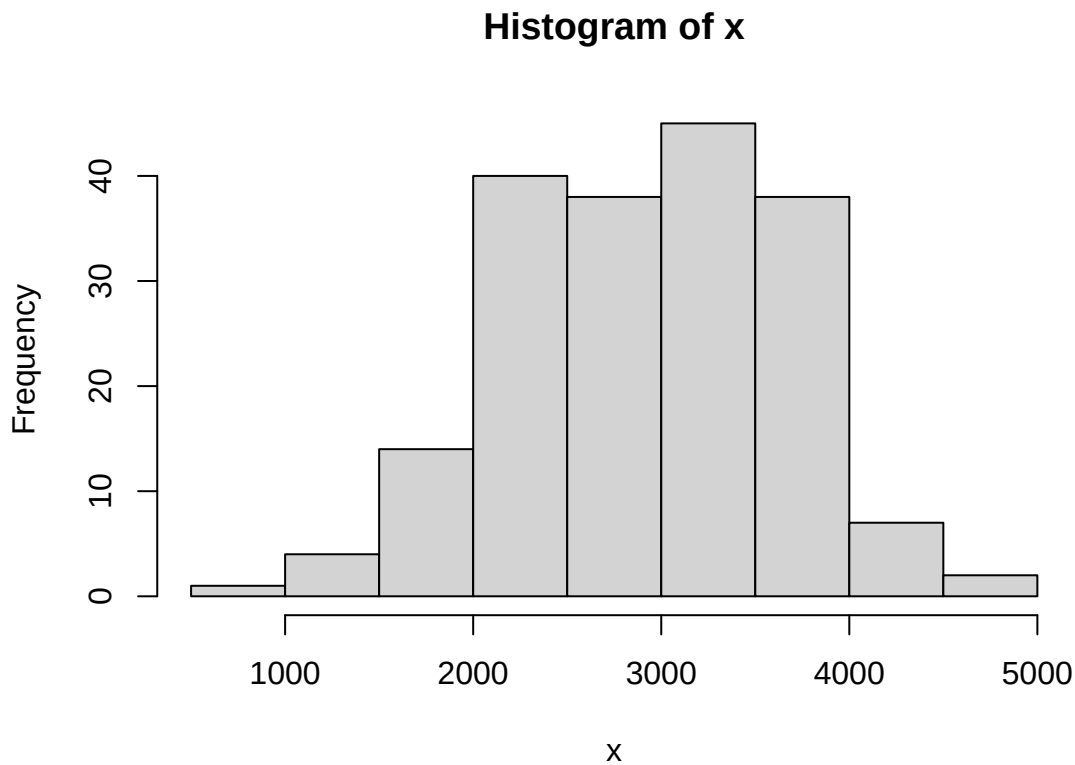
FALL 2024

HISTOGRAMS

The following data represent the birthweights (in grammes) of $n = 189$ newborns.

```
## [1] 2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 2722 2733 2751 2750
## [15] 2769 2769 2778 2782 2807 2821 2835 2835 2836 2863 2877 2877 2906 2920
## [29] 2920 2920 2920 2948 2948 2977 2977 2977 2977 2922 3005 3033 3042 3062
## [43] 3062 3062 3062 3062 3080 3090 3090 3090 3100 3104 3132 3147 3175 3175
## [57] 3203 3203 3203 3225 3225 3232 3232 3234 3260 3274 3274 3303 3317 3317
## [71] 3317 3321 3331 3374 3374 3402 3416 3430 3444 3459 3460 3473 3544 3487
## [85] 3544 3572 3572 3586 3600 3614 3614 3629 3629 3637 3643 3651 3651 3651
## [99] 3651 3699 3728 3756 3770 3770 3770 3790 3799 3827 3856 3860 3860 3884
## [113] 3884 3912 3940 3941 3941 3969 3983 3997 3997 4054 4054 4111 4153 4167
## [127] 4174 4238 4593 4990 709 1021 1135 1330 1474 1588 1588 1701 1729 1790
## [141] 1818 1885 1893 1899 1928 1928 1928 1936 1970 2055 2055 2082 2084 2084
## [155] 2100 2125 2126 2187 2187 2211 2225 2240 2240 2282 2296 2296 2301 2325
## [169] 2353 2353 2367 2381 2381 2381 2410 2410 2410 2414 2424 2438 2442 2450
## [183] 2466 2466 2466 2495 2495 2495 2495
```

We can plot a simple histogram: the default plot in R looks like this.



We can see from the plot that there are $k = 9$ intervals, and that the interval end points are

500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000

but we cannot easily see the individual frequencies for each category; they are

1, 4, 14, 40, 38, 45, 38, 7, 2.

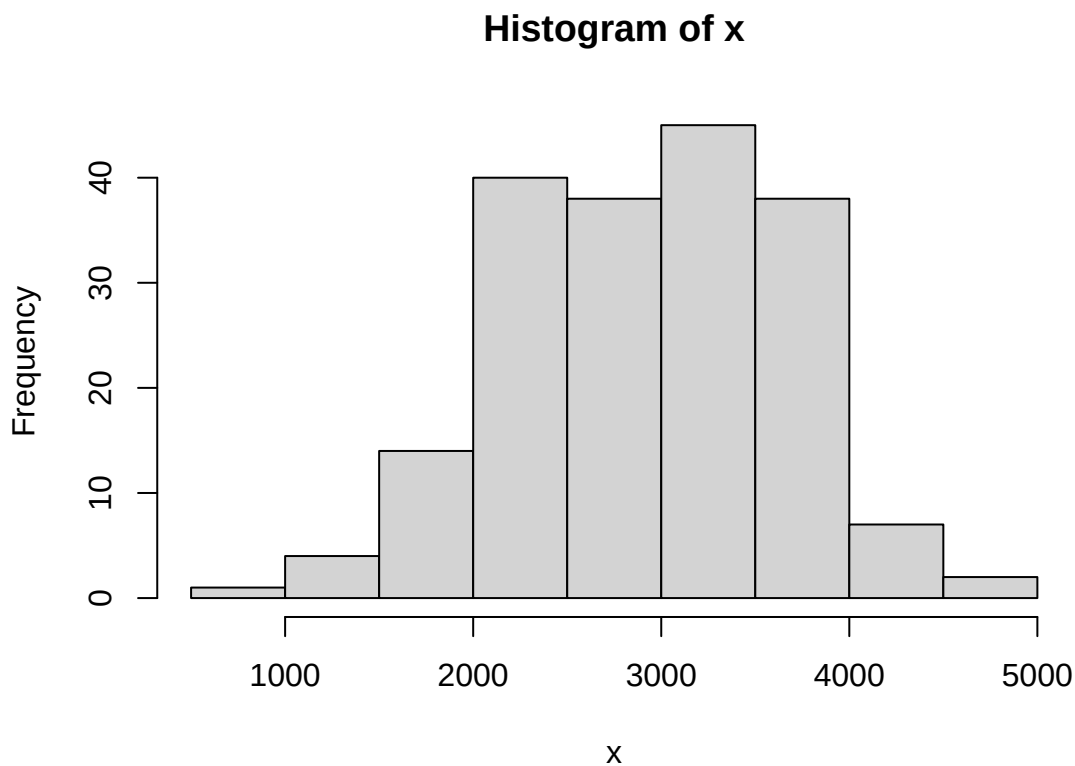
The default plot in R does not use the categorization rule as described in lectures, that is, the default rule in R puts a data value x into an interval with endpoints a and b if

$$a < x \leq b$$

whereas the method outlined in lectures uses the rule

$$a \leq x < b.$$

Here, however, if we use the rule from lectures, we get



We have the individual frequencies for each category

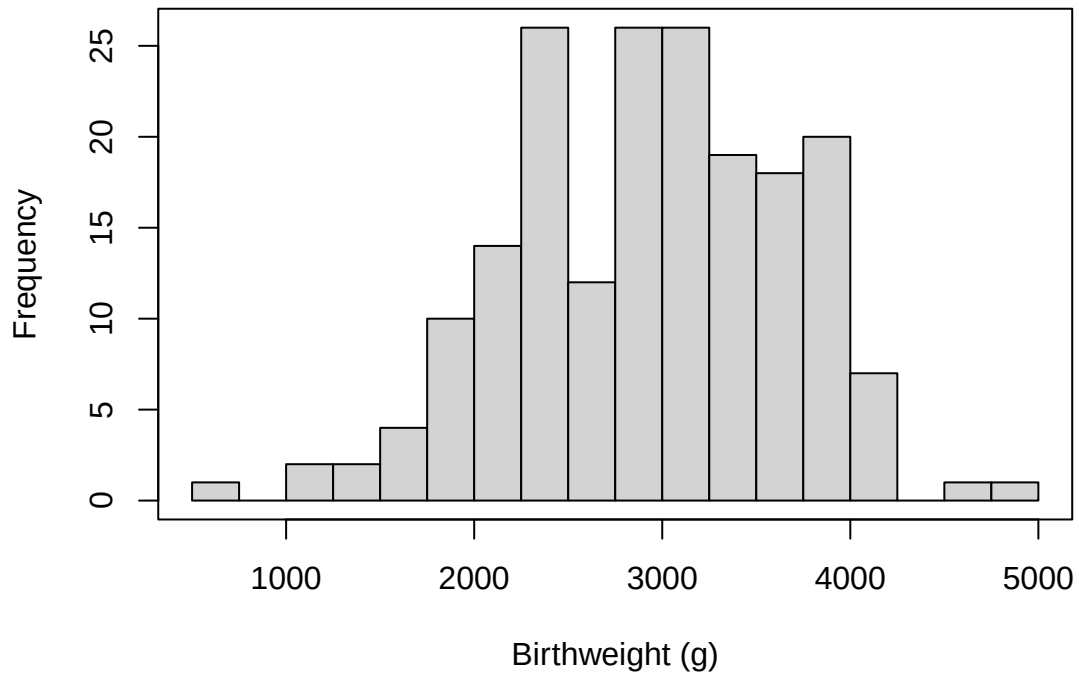
1, 4, 14, 40, 38, 45, 38, 7, 2

and there is no difference, as no individual data point matches any of the interval end points.

We can add plotting options to make the plot more interpretable. In the next plot the following aspects have been altered:

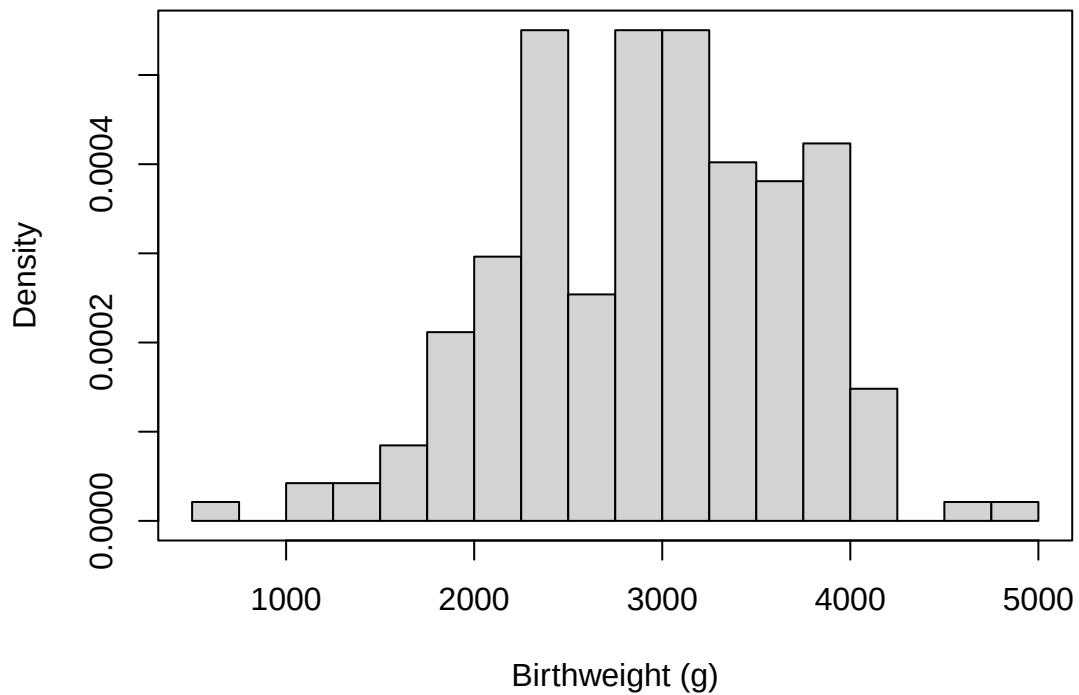
- the number of intervals is increased to 18 covering the same range,
- the x -axis is re-labelled to be more informative,
- an informative title has been added.

Birthweight data for n=189 newborns



In the next plot, the same data are used in a **density** plot:

Birthweight data for n=189 newborns



Note the change in the vertical scale: the frequency plot uses the frequencies

1, 0, 2, 2, 4, 10, 14, 26, 12, 26, 26, 19, 18, 20, 7, 0, 1, 1

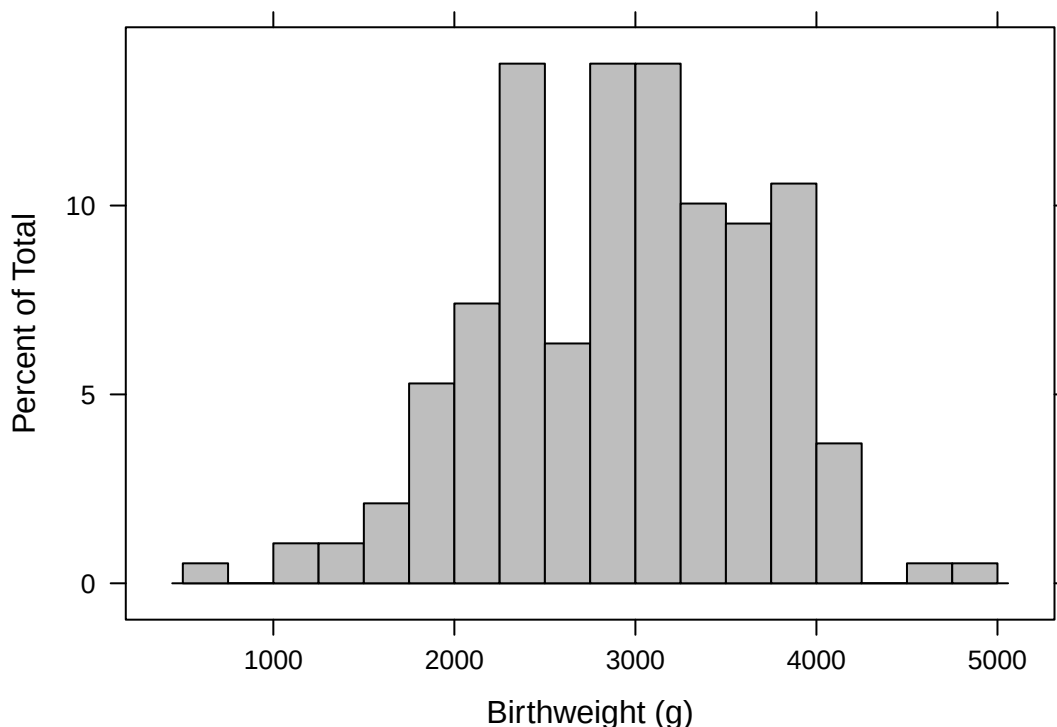
to determine the height of each rectangle, whereas the density version takes a relative frequency, and divides it by the interval width (the 'density' is measured per unit of the x -axis). In this plot the interval width is 250, so the density value for the eighth interval is

$$\frac{\text{Frequency in eighth interval}}{\text{Total sample size} \times \text{Interval width}} = \frac{26}{189 \times 250} = 0.0006$$

In this case, there is no difference in the appearance of the plot, simply a change in the vertical axis values. This is because all the interval widths are the same.

We can also plot a **relative frequency** plot:

Birthweight data for $n=189$ newborns



Here the relative frequencies are the frequencies per interval divided by the sample size. The relative frequencies are interval frequencies listed above divided by $n = 189$ (here rounded to 4 decimal places):

```
## [1] 0.0053 0.0000 0.0106 0.0106 0.0212 0.0529 0.0741 0.1376 0.0635 0.1376
## [11] 0.1376 0.1005 0.0952 0.1058 0.0370 0.0000 0.0053 0.0053
```

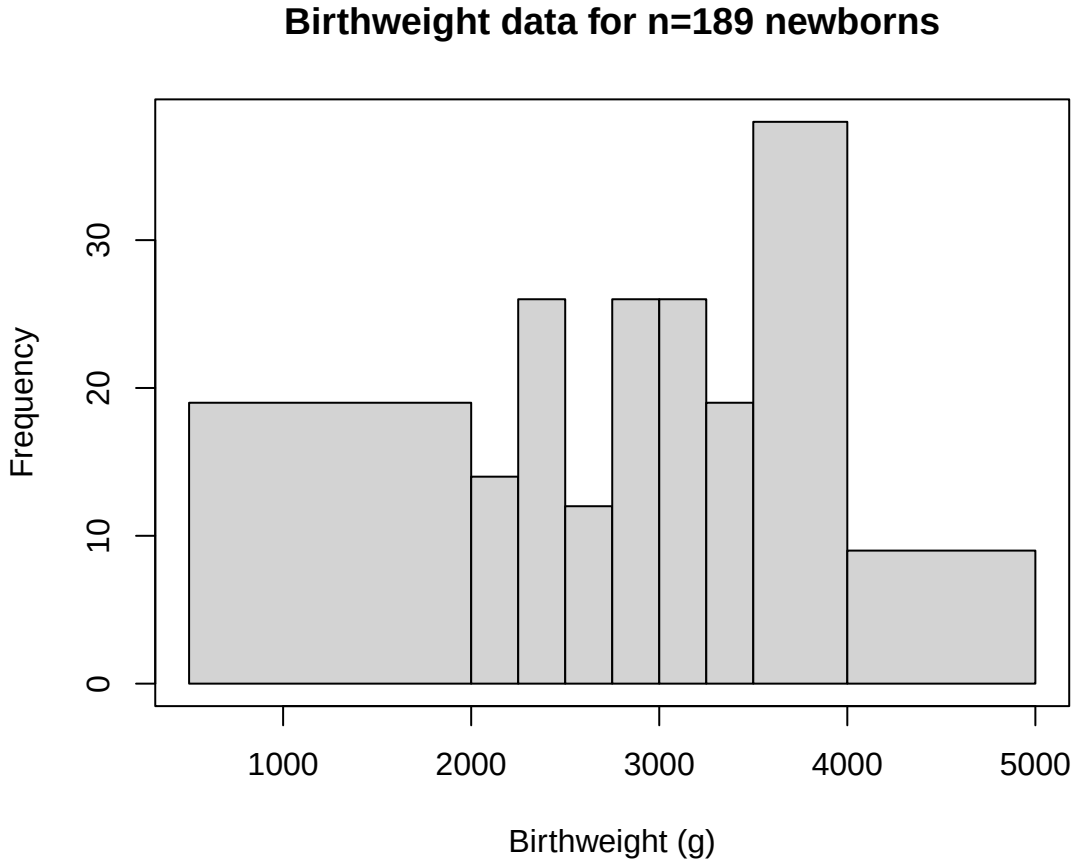
In the plot, these relative frequencies are plotted as percentages (i.e. multiplied by 100).

```
## [1] 0.5291 0.0000 1.0582 1.0582 2.1164 5.2910 7.4074 13.7566
## [9] 6.3492 13.7566 13.7566 10.0529 9.5238 10.5820 3.7037 0.0000
## [17] 0.5291 0.5291
```

It is possible to use unequal-sized intervals, but this makes interpretation of the histogram slightly more complicated. Suppose we want to use the intervals

$[500, 2000)$, $[2000, 2250)$, $[2250, 2500)$, $[2500, 2750)$, $[2750, 3000)$, $[3250, 3500)$, $[3500, 4000)$, $[4000, 5000]$

that is with the first interval width 1500g, the eighth interval width being 500g, the last interval width being 1000g, and the other interval widths being 250g. The corresponding histogram is



There are $k = 9$ intervals and the individual frequencies for each of the intervals are

19, 14, 26, 12, 26, 26, 19, 38, 9.

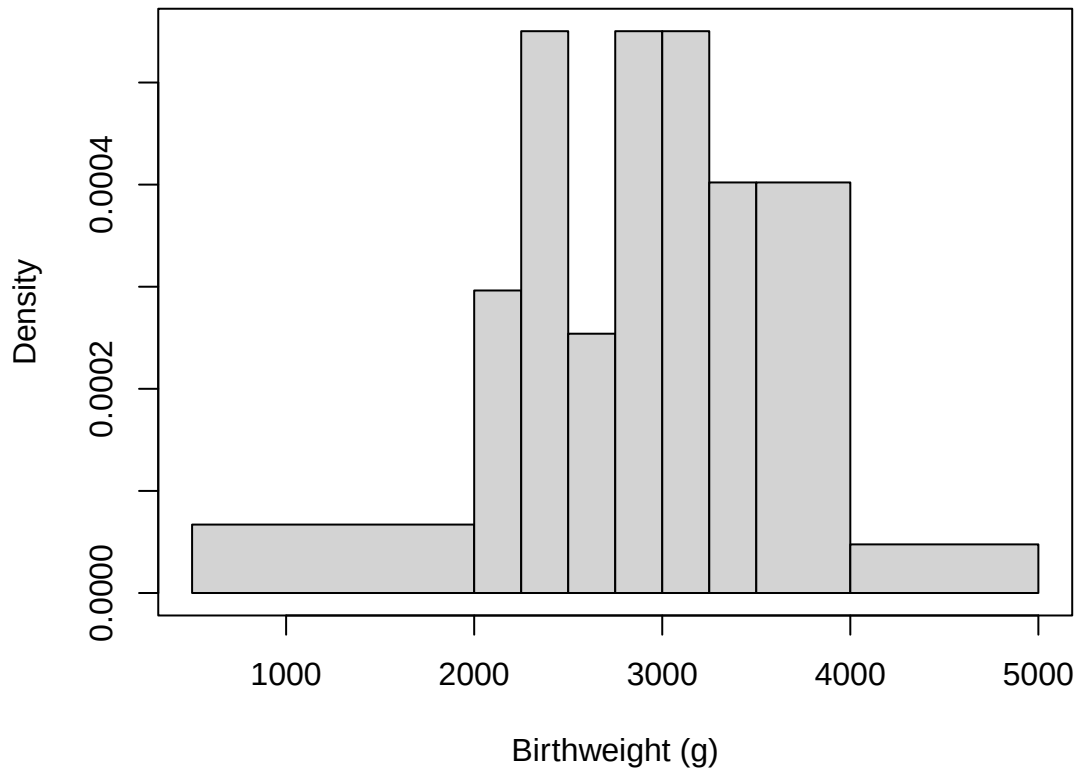
Due to the fact that the intervals are unequal sizes, it is preferable to use a density plot: here the density value for each interval is computed using the interval width for that interval. For example, for the first interval, we have

$$\frac{\text{Frequency in first interval}}{\text{Total sample size} \times \text{Interval width for first interval}} = \frac{19}{189 \times 1500} = 0.00007$$

whereas for the fifth interval, we have

$$\frac{\text{Frequency in fifth interval}}{\text{Total sample size} \times \text{Interval width for fifth interval}} = \frac{26}{189 \times 250} = 0.00055$$

Birthweight data for n=189 newborns



SUMMARY

Key elements of the histogram:

1. **Number of intervals:** how many rectangles are used ?

2. **Interval boundaries:**

- do the rectangles have equal widths or unequal widths ?
- which of the rules

$$a \leq x < b \quad \text{or} \quad a < x \leq b$$

is being used to categorize data points lying exactly on interval boundaries ?

3. **Horizontal scale and label:** what is the variable of interest and its measurement unit ?

4. **Vertical scale:** what type of vertical scale is being used

- frequency ?
- relative frequency ?
- density ?

5. **Title.**