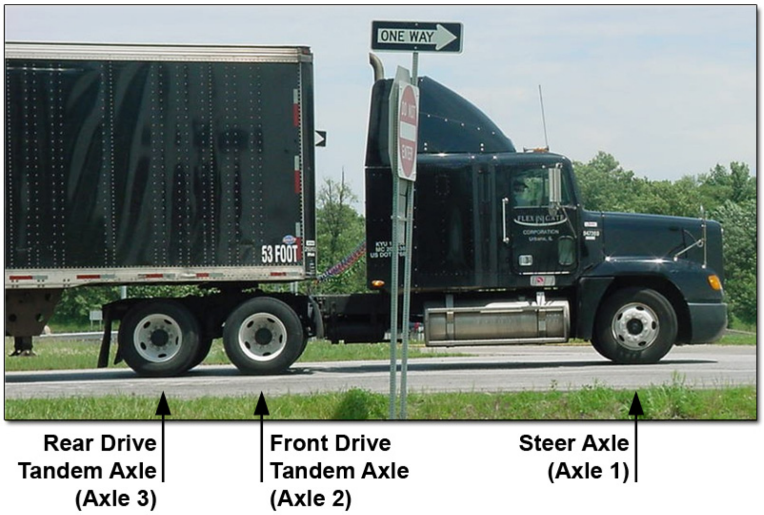
Kernel Density Estimates and Histograms

ACTIVITY 29

Figure 50 Axles 1, 2 and 3 (Picture: Andrew Nichols)



The histogram and the kernel density estimate are two great tools that allow you to quickly see the distribution of a continuous variable. You are probably familiar with the histogram, but probably not with its nuances. We'll explore those. The kernel density estimate (KDE) might be new to you but you will grow to love it. We'll start by looking at the WIM data and then you will get a chance to explore further. We will focus on the axle weights of 5-axle semi-trucks in this exploration.

Purpose

To combine lessons from descriptive statistics with an exploration of data.

Learning Objective

Use graphical analysis and scripting to identify possible data quality issues.

Required Resources

- R, R Studio
- PostgreSQL ODBC driver
- Direct connection to PostgreSQL with R
- Sample R script

Time Allocated

90 minutes in class

Tasks

Open the R script for Activity 29 from the class website. We will start our exploration with the just the class 11 trucks from all stations. Note that we have sped up the read from the class database if you only select the columns you need, rather than using the all (*) SQL operator.

```
qry <- " SELECT stationnum, axl1, axl2,axl3, axl4, axl5 FROM wim.wimdata
WHERE type='11'"
```

Later, you will be analyzing some 750,258 rows of data. For parts A and B of this activity, we will subset the data to just station 8.

A. Introduction to the Histogram

The main display decision for the histogram is how many bins to count the frequency of the data. The number of bins can be set with the `breaks=` option in the call to histogram. In Keen, the bin width is the class width. The class width is evenly divided over the range of the data (rounded to an integer). Keen lists and describes ways to determine the bin width and number of classes:

- Rule of Twelve
- Robust Rule of Twelve

- Sturges Rule
- Doanes Rule
- Scott’s Rule
- Freedman-Diaconis Rule

In R, it can be hard-coded with an integer, or specified in a call to a function. Options are Sturges, Scott and Freedman-Diaconis. Run the code that creates the histograms shown in Figure 51. Note that the y-axis, which is the number in each bin, changes for each.

1. Create the same set of histograms for axl2, axl3, axl4, axl5. Can you see the bimodal aspect of any of the distributions? Is it easier to see with one of the break settings?

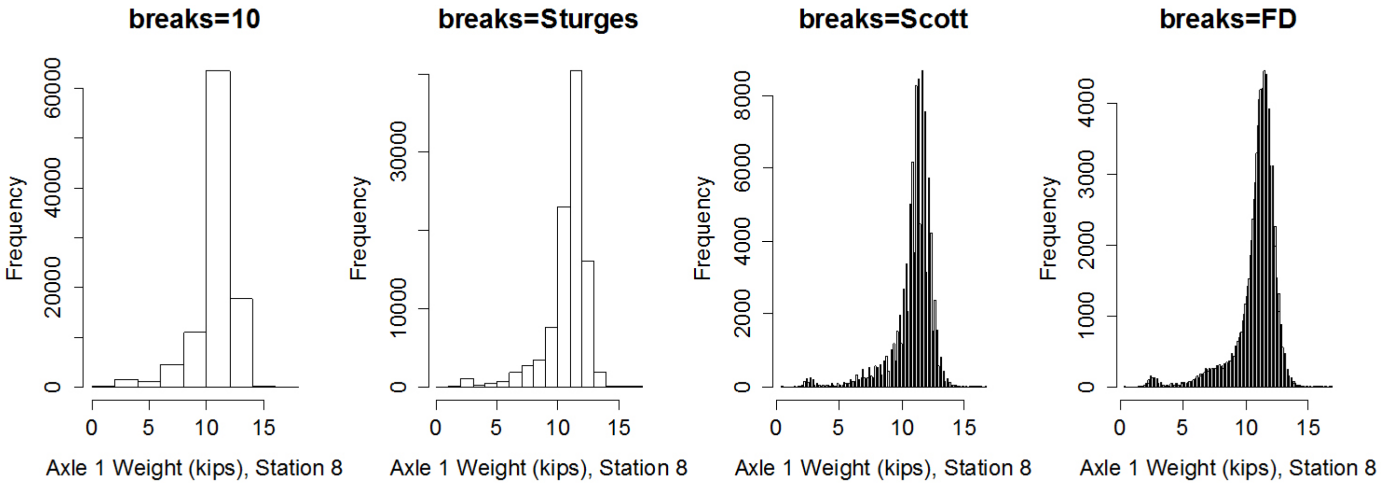


Figure 51 Histograms of Station 8, Axle 1 Weight in Kips, August 2009

Another useful option is that you can store the results of the histogram without plotting them by setting the call to plot=FALSE. Then you extract the names count, density, or breaks vectors.

```
y <- hist (wim8$axl1, breaks=10, plot=FALSE)
str(y) # see the values of the histogram class
y$breaks # pull out the vector of breaks
y$counts # set the counts
sum(y$density)  #what should this sum to ?
```

You can also plot the histogram with the density (#bin/total obs) instead of the frequency on the y-axis. With this call:

```
hist (wim8$axl1, breaks=”Sturges”, freq=FALSE)
```

B. Introduction to Kernel Density Estimate (KDE)

The kernel density estimate plot is essentially a smooth histogram. The resulting plot is non-parametric estimate of the variable under consideration’s density function. It is akin to the pdf of random variable. You have a choice of the bandwidth and kernel. The bandwith is the effective width of the sliding window used to generate the density. As the sample gets larger, the bandwidth can be made larger. The kernel is function over which the density is estimated. It can be one of the following in R “gaussian”, “epanechnikov”, “rectangular”,”triangular”, “biweight”,”cosine”, “optcosine”).

The density plot is generated with a call to density function first, which can then be plotted:

```
denmass <- density(wim8$axl1, kernel="gaussian")
plot(denmass, main="gaussian")
```

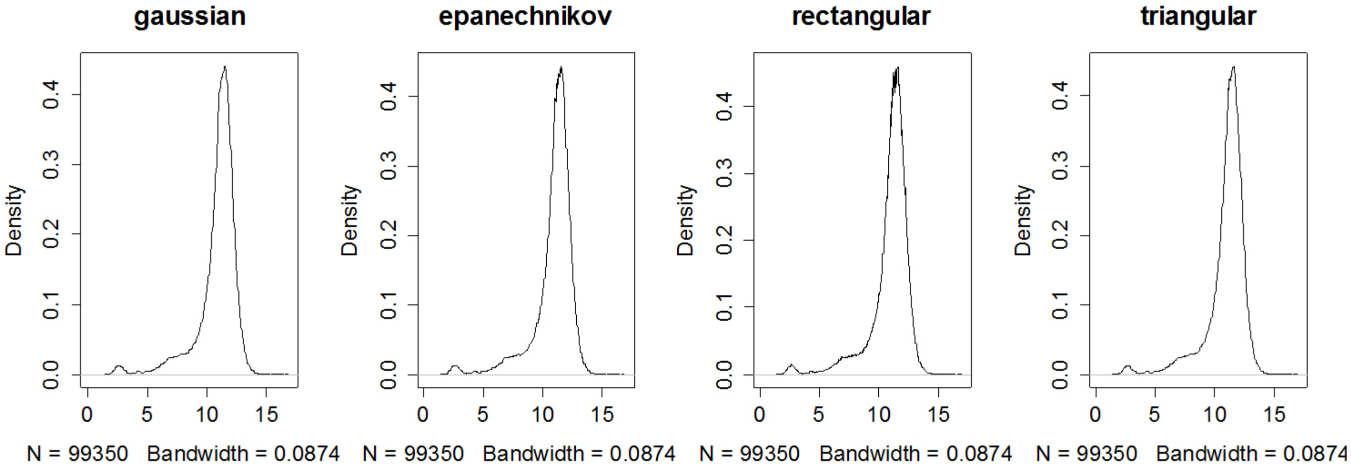


Figure 52 KDE plots of Station 8, Axle 1 Weight in Kips, August 2009

Modify the KDE to explore the effect of the `bw=setting`. Check the function `bw.nrd0()` which returns the default bandwidth. You should see how large bandwidths smooth out the KDE.

C. On Your Own

The steering axle weight of the 5-axle semi truck (in the *wim* data type 11 trucks, Class 9 in the FHWA scheme) ranges between 9,000-11,000 lbs. This is the steering axle and mainly handles the weight of the engine and tractor. If there is a lot variability in these data or the observations of one station are lower than the others, it could be an indication that a station is not performing correctly and that calibration may be required.

Using the KDE plots, explore the station to station variability of axle 1 weight to show the distributions of the axle1 weight for all stations. In these plots, look for different shapes of the distribution and for differences in the in the central location of the distribution. First, set up the plots so that they are generated in their own frame and arrange the plots. Then, try to put KDEs of the plots in all one plot on the same scale (using the lines function). You should also use other plots that we have explored (such as the boxplot) to help you diagnose variability and median means.

Repeat this same exercise for the other axle weights. With these weights, it is harder to see differences, since we have the effect of loaded and unloaded trucks.

Deliverable

Prepare a short write up of your discovery about the underperforming stations. In your write-up, clearly identify the stations that appear to be in need of calibration or error checking. Elaborate on your reasons. Submit a PDF of this write up to the drop box. Include R code as an appendix.

Assessment

This activity is a short response activity. The score that you receive will be based on the quality and depth of discussion. The response expected differs by question as described in rubric below:

Activity 29 Grading Rubric

	Excellent (10)	Good (8)	Poor (6)	NONE
Discussion	Insightful discussion or commentary relating to the question at hand demonstrating student understanding of the task.	Discussion was competent regarding the lessons, but lacked in insightful discussion.	The discussion did not address the lessons or applicability of the activity.	Did not submit
Graphic(s)	Graphic correctly demonstrated the distribution of the data set.	Graphic did not accurately demonstrated the distribution of the data.	No graphics were present in the discussion.	Did not submit
Quality	Document is typed, formatted, contains appropriate grammar and language.	Document has minor grammatical errors or inappropriate language.	Document was unorganized, contained inappropriate language and/or grammatical errors.	Did not submit